Landmark Selection for Zero-shot Learning

Yuchen Guo[1], **Guiguang Ding**[2], **Jungong Han**[3], **Chenggang Yan**[4], **Jiyong Zhang**[4] and **Qionghai Dai**[1]

[1]Department of Automation, Tsinghua University, Beijing, China
[2]School of Software, Tsinghua University, Beijing, China
[3]WMG Data Science, University of Warwick, Coventry, UK
[4]School of Automation, Hangzhou Dianzi University, China

Abstract

Zero-shot learning (ZSL) is an emerging research topic whose goal is to build recognition models for previously unseen classes. The basic idea of ZSL is based on heterogeneous feature matching which learns a compatibility function between image and class features using seen classes. The function is constructed based on one-vs-all training in which each class has only one class feature and many image features. Existing ZSL works mostly treat all image features equivalently. However, in this paper we argue that it is more reasonable to use some representative cross-domain data instead of all. Motivated by this idea, we propose a novel approach, termed as **La**ndmark **S**elec**t**ion(LAST) for ZSL. LAST is able to identify representative cross-domain features which further lead to better image-class compatibility function. Experiments on several ZSL datasets including ImageNet demonstrate the superiority of LAST to the state-of-the-arts.

1 Introduction

When training a recognition model for a category, it is expected to have sufficient labeled samples for training in a standard supervised learning way. However, this requirement is too demanding in many real-world applications, such as Web image classification and fine-grained classification. In these scenarios, there are a large number of classes, in which some common ones have sufficient labeled examples while most uncommon ones have only few or no labeled data, since the number of labeled samples follows a long-tail distribution [Changpinyo *et al.*, 2016]. To deal with this challenge, zero-shot learning (ZSL) [Farhadi *et al.*, 2009; Lampert *et al.*, 2014] has been demonstrated to be a promising solution. ZSL learns a compatibility function between image and class features by the data from seen classes. Then this function is transferred to the unseen ones for prediction.

ZSL is currently a hot research topic attracting considerable research interest. Many approaches have been proposed yielding promising performance [Xian *et al.*, 2017]. Formally, most of representative approaches are formulated as a heterogeneous domain adaptation problem which aligns the image feature space and class feature space, such as the deep networks' output [He *et al.*, 2016] and the class attributes [Lampert *et al.*, 2014] or word2vec embeddings [Huang *et al.*, 2012]. The alignment function, or the compatibility function, is learned from the seen classes and their labeled samples. By assuming the transferability of the function between classes, one can apply it to unseen classes even though they do not have labeled images for training.

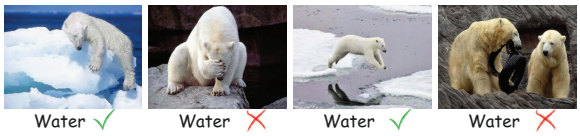


Figure 1: The "polar bear" class has the attribute "water" in AwA dataset, while not all images contain it. It is clearly unreasonable to align all image features indiscriminately to the class attribute vector.

How to learn the compatibility function is the key problem in ZSL literatures [Xian *et al.*, 2017]. Based on the properties of data, one-vs-all training is always utilized. In particular, each class has one class feature vector and many images. Therefore, the class vector is combined with each of the image vectors to compute the loss function, which indicates that all images contribute equally to the training procedure.

It is necessary to raise a question: is every image equally important for training? The answer has been given by many literatures, including active learning [Aggarwal *et al.*, 2014] and hard example mining [Shrivastava *et al.*, 2016], which demonstrates that using the most representative and informative samples may leads to better performance. In the scenario of ZSL, it is more reasonable to assign different weights to different images. In fact, each class has only one feature vector, like the attribute vector. This vector usually indicates the general characteristics of the class. On the other hand, different images may show different characteristics of a class due to the various occlusion, object size, orientation, and so on. For example, a "polar bear" class has the attribute "water". However, this attribute does not appear in all "polar bear" images since some images are captured on the ground. In this circumstance, it seems unreasonable to match a "non-water" image to the "water" attribute. Moreover, due to the missing attributes, the image-class pair yields large training loss. To minimize the total loss, the model will try to fit these mismatching pairs, resulting a biased model from a good one.

Here we show an illustration in Figure 1. The Animals with Attributes (AwA) [Lampert *et al.*, 2014] uses an attribute vector for each class. The "polar bear" has the attribute "water". However, many "polar bear" images do not contain this attribute. Obviously not all "polar bear" images are representative for the class, especially when described by only one attribute vector. To learn a good compatibility function between class features and image features in the one-vs-all setting, it is necessary to find the most representative images which are the most compatible with the class feature for model training.

Based on this motivation, in this paper we propose a novel approach, **La**ndmark **S**elec**t**ion (LAST) for ZSL. Instead of regarding all images equally important for compatibility function learning, LAST exploits the representative images which are then associated to the class features. We jointly optimize the landmark selection and function learning from a cold start, which turns out to be an effective strategy. In addition, different from most existing ZSL approaches which use sample-wise or class-wise matching, we further introduce a global domain alignment term into the objective function, which is inspired by domain adaptation [Pan and Yang, 2010; Tsai *et al.*, 2016]. Based on this global alignment term, the model is capable of generalizing better for unseen classes and images, which is desired for ZSL. Both global term and class-specific conditional term are formulated into a joint optimization function. It improves the performance of ZSL, especially for large-scale benchmarks which have many classes. In summary, we make the following contributions in this paper.

1. We notice that it is unreasonable to assign the same weight to all images during training a ZSL model in one-vs-all strategy. To address this issue, we propose LAST which finds representative images to match the class features. The selected landmarks result in more effective ZSL models.

2. Inspired by domain adaptation, we further incorporate a global alignment term into the objective function together with the conditional term, making the model generalize better, especially for large-scale benchmarks with many classes.

3. We carry out extensive experiments on several benchmarks, including ImageNet [Russakovsky *et al.*, 2015]. The results demonstrate the effectiveness of LAST for (generalized) ZSL in comparison with the state-of-the-art approaches.

2 Preliminary and Related Work

2.1 Notations

Following the definition in [Xian *et al.*, 2017], we describe ZSL problem as follows. There are two disjoint class sets $\mathcal{C}_s = \{c_1^s, ..., c_{k_s}^s\}$ and $\mathcal{C}_u = \{c_1^u, ..., c_{k_u}^u\}$ with $\mathcal{C}_s \cap \mathcal{C}_u = \emptyset$, denoted as seen classes and unseen classes respectively. Each image is represented by an image feature vector $x \in \mathbb{R}^p$ and each class is represented by a label feature vector $y \in \mathbb{R}^q$. There is a training set $\mathcal{D}_{tr} = \{(x_i, y_i)\}_{i=1}^{n_s}$ where the each class feature y_i corresponds to a seen class from $\mathcal{C}_s$. A compatibility function $F(x, y; W)$ between image and class features is trained based on the training set. Then, it is applied to a test sample, which is from unseen classes $\mathcal{C}_u$ in the conventional ZSL setting, or $\mathcal{C}_s \cup \mathcal{C}_u$ in the generalized ZSL setting. The classification is performed by selecting the class which has the largest compatibility to the test sample by $F(x, y; W)$.

2.2 Zero-shot Learning

The key component in ZSL is the compatibility function $F(x, y; W)$. Given an unseen class, we can use F to compute its relationship to any images even though no visual example of this class is available for training. Most ZSL approaches focus on the learning algorithm and the specific form for F.

One widely used definition is the bilinear model $F(x, y; W) = xWy'$. It has been used in Attribute Label Embedding (ALE) [Akata *et al.*, 2016], Deep Visual Semantic Embedding (DEVISE) [Frome *et al.*, 2013], Structured Joint Embedding (SJE) [Akata *et al.*, 2015], Embarrassingly Simple ZSL (ESZSL) [Romera-Paredes and Torr, 2015], and so on [Guo *et al.*, 2016; Zhang and Saligrama, 2016]. To train the model, different loss functions are utilized, including ranking loss, triplet loss, Euclidean loss, cross-entropy loss, and etc. Besides, some other compatibility functions are equivalent to the bilinear function. For example, Euclidean distance based function, like $F(x, y; W) = -\|xW_x - y\|_2$ or $F(x, y; W) = -\|x - yW_y\|_2$ [Kodirov *et al.*, 2017]. Latent Embedding (LATEM) [Xian *et al.*, 2016] uses multiple compatibility function with latent variables $F(x, y; W) = \max_{1 \le i \le K} xW_i y$. Moreover, some approaches try to learn image-specific feature transformation and/or class-specific feature transformation to improve the non-linearity of the function. Cross-modal Transfer (CMT) [Socher *et al.*, 2013] utilizes $\tanh(xW_x)$ for image feature transformation. Semantic Similarity Embedding (SSE) [Zhang and Saligrama, 2015] uses sparse coding for image transformation and ReLU for class transformation. Implicit Non-linear Similarity Scoring (ICINESS) [Guo *et al.*, 2018] uses deep convolutional networks for image transformation and multi-layer perceptron for class transformation. Discriminative Semantic Representation Learning (DSRL) [Ye and Guo, 2017] uses sparse non-negative matrix factorization and max-margin semantic alignment for class transformation. Though they have different details, they basically follow the bilinear model.

Besides, some ZSL approaches use other ideas. For example, sample transfer [Guo *et al.*, 2017b] and sample synthesis [Guo *et al.*, 2017a] focus on bring in pseudo examples for unseen classes. Direct/Indirect Attribute Prediction (DAP/IAP) [Farhadi *et al.*, 2009; Lampert *et al.*, 2014] consider to recognize the attributes from images and compare them to the class attributes. Convex Semantic Embedding (CONSE) [Norouzi *et al.*, 2013] firstly compute the probability of an image belonging to a seen class. Then based on the similarity between class features, the probability is propagated to the unseen classes. Semantic Manifold Distance (SMD) [Fu *et al.*, 2015] further utilizes a similarity graph constructed from class features and hidden Markov process to propagate the probability. Synthesized Classifier (SYNC) [Changpinyo *et al.*, 2016] and Shared Model Space (SMS) [Guo *et al.*, 2016] learn a mapping from class features to the classifier parameter space. Semantics-Preserving Adversarial Embedding Networks (SP-AEN) [Chen *et al.*, 2018] uses generative adversarial networks to construct visual examples for the unseen classes. Cycle-consistent Generalized ZSL [Felix *et al.*, 2018] considers to adopt multi-modal cycle-consistent GAN to generate image representations.

3 The Proposed Approach

3.1 Objective Function

In this paper, we regard ZSL as a domain adaptation problem, which learns a transformation matrix W to align image feature x and class feature y. We base our approach on the ideas of autoencoder [Kodirov *et al.*, 2017] and cycle-GAN [Felix *et al.*, 2018]. The base objective function is:

$$\min_{W} \sum_{i=1}^{n_s} \|x_i W - y_i\|_2^2 + \lambda \|x_i - y_i W'\|_2^2 + \gamma \|W\|_F^2 \quad (1)$$

We hope the compatibility matrix W can well align heterogeneous features such that our objective is to minimize the distortion of both image-to-class projection and class-to-image projection. This function can be regarded as the base model.

Obviously, Eq. (1) is class aware and it connects image features and class features belonging to the same class. Therefore, we can regard Eq. (1) as minimizing the difference between the conditional distribution of image features and class features. Inspired by the basic idea of domain adaptation [Pan and Yang, 2010], we propose to further consider the marginal distribution difference between image features and class features. In fact, one goal of learning the compatibility matrix W is to align the image feature space and class feature space. Based on this motivation, we propose to further minimize the maximum mean discrepancy (MMD) [Wang and Deng, 2018] between the projected image features and the projected class features as follows,

$$\begin{aligned} &\min_{W} \| \frac{1}{n_s} \sum_{i=1}^{n_s} x_i W - \frac{1}{k_s} \sum_{c=1}^{k_s} y^c \|_2^2 \\ &+ \beta \| \frac{1}{n_s} \sum_{i=1}^{n_s} x_i - \frac{1}{k_s} \sum_{c=1}^{k_s} y^c W' \| + \gamma \|W\|_F^2 \end{aligned} \quad (2)$$

If Eq. (1) is regarded as the class-specific alignment, Eq. (2) can be regarded as the global alignment. By reviewing existing ZSL literatures, it is observed that Eq. (2) is always ignored. In fact, ZSL focuses more on the unseen classes instead of the seen ones. Therefore, when using seen classes for training, the generalization ability of the model is an important issue. To improve the generalization ability, it seems reasonable to consider the global alignment between heterogeneous feature spaces. In the experiments, we will show the effectiveness of this term. By combining Eq. (1) and Eq. (2), we obtain an objective function based on domain adaptation.

As discussed above, Eq. (1) and many other ZSL approaches adopt a one-vs-all strategy where each of the k_s class features is combined with all image features belonging to this class. In this way, all images belonging to this class have the same training weight. However, not all of them are representative for this class, as illustrated in Figure 1. To improve the performance, it seems reasonable to identify the adaptation ability of each image with a proper weight. These images are the landmarks for adaptation which indicates that they are the most representative to reflect the properties (attributes) of the class like the class feature. As introduced, each class has only one class feature, which shows the most general characteristics of this class. Therefore, it is also desired that the selected images could contain general characteristics such that they can be correctly aligned to the class feature. To do so, we introduce a learnable non-negative weight μ_i for each image. Taking the weight for each image into account, the objective function of the proposed LAST is:

$$\begin{aligned} \min_{W,\mu_i} &\sum_{i=1}^{n_s} \mu_i \|x_i W - y_i\|_2^2 + \lambda \sum_{i=1}^{n_s} \mu_i \|x_i - y_i W'\|_2^2 \\ &+ \alpha \| \frac{1}{\delta n_s} \sum_{i=1}^{n_s} \mu_i x_i W - \frac{1}{k_s} \sum_{c=1}^{k_s} y^c \|_2^2 \\ &+ \beta \| \frac{1}{\delta n_s} \sum_{i=1}^{n_s} \mu_i x_i - \frac{1}{k_s} \sum_{c=1}^{k_s} y^c W' \|_2^2 + \gamma \|W\|_F^2 \\ s.t.\ &\mu_i \in [0,1],\ \sum_{i=1}^{n_s} \mu_i = \delta n_s \end{aligned} \quad (3)$$

where the first two terms consider the conditional alignment, the third and the fourth terms consider the marginal alignment, and the fifth term is a regularization to control the complexity. The weight parameter μ_i indicates if x_i a representative image. A non-zero weight means that this image is selected as a landmark. Theoretically, using 0-1 binary weight seems more compatible with "landmark selection". However, binary weight makes the optimization complicated. Therefore we use continuous value in $[0, 1]$ as soft selection. This value can reflects the importance of an image and its contribution to the alignment. The parameter $\delta \in [0, 1]$ determines the portion of the selected landmarks. For example, we can set $\delta = 0.5$ to softly use half of images.

It is easy to observe the difference between LAST and previous ZSL approaches. Firstly, unlike the other approaches regarding all images equally important, we propose to use (soft) landmark selection to find the representative images for cross-domain adaptation. Since each class has only one class feature, it is more reasonable to use the representative images instead of the ones with micro and specific characteristics. Secondly, for domain adaptation, LAST not only considers the conditional alignment like previous approaches, but it also utilizes the marginal alignment between image feature space and class feature space. This strategy is capable of aligning two heterogeneous spaces from a global perspective, making the model generalize better for previously unseen classes.

3.2 Optimization

Eq. (3) has two variables to optimize, the compatibility matrix W and the selection weight μ_i. One simple strategy is to use an iterative optimization algorithm which alternatively updates one of them while keeping the other variable fixed.

Optimizing $\mathbf{W}$. Denote $\Omega = diag(\mu_1, ..., \mu_{n_s})$, $\mu = [\mu_1, ..., \mu_{n_s}]$ and fix μ. We can rewrite Eq. (3) as follows,

$$\begin{aligned} \min_{W} &\|(\Omega X)W - \Omega Y\|_F^2 + \lambda \|\Omega X - (\Omega Y)W'\|_F^2 \\ &+ \alpha \| \frac{1}{\delta n_s} (\mu X)W - \frac{1}{k_s} e_{k_s} Y_C \|_2^2 \\ &+ \beta \| \frac{1}{\delta n_s} \mu X - \frac{1}{k_s} (e_{k_s} Y_C) W' \|_2^2 + \gamma \|W\|_F^2 \end{aligned} \quad (4)$$

where e_{k_s} is a k_s-long row vector whose all elements are 1, $X = [x_1; ...; x_{n_s}] \in \mathbb{R}^{n_s \times p}$, $Y = [y_1; ...; y_{n_s}] \in \mathbb{R}^{n_s \times q}$, and $Y_C = [y^1; , , , ; y^{k_s}] \in \mathbb{R}^{k_s \times q}$. Now we need to take a derivative of the loss function above with respect to W and set it to zero to solve W, which leads to the problem below,

$$AW + WB = C \tag{5}$$

where

$$A = X'\Omega'\Omega X + \frac{\alpha}{\delta^2 n_s^2} X'\mu'\mu X + \gamma$$

$$B = \lambda Y'\Omega'\Omega Y + \frac{\beta}{k_s^2} Y_C' e_{k_s}' e_{k_s} Y_C$$

$$C = (1+\lambda) X'\Omega'\Omega Y + \frac{\alpha+\beta}{\delta n_s k_s} X'\mu' e_{k_s} Y_C$$

Eq. (4) is a Sylvester equation which can be efficiently solved by the Bartels-Stewart algorithm [Bartels and Stewart, 1972]. For example, it can be solved in MATLAB by a build-in function $\mathrm{sylvester}$[1]. The complexity of the solver depends on the dimensionality of features, and is independent of the number of samples, making it feasible for large-scale datasets.

Optimizing μ. By fixing W, we can rewrite Eq. (3) as

$$\min_{\mu_i \in [0,1], \mu e_{n_s}' = \delta n_s} \mu F \mu' + \mu g \tag{6}$$

where

$$F = \frac{\alpha}{\delta^2 n_s^2} XWW'X' + \frac{\beta}{\delta^2 n_s^2} XX'$$

$$g = r_x + \lambda r_y - \frac{2(\alpha+\beta)}{\delta n_s k_s} XWY_C' e_{k_s}'$$

$$r_x = diag((XW - Y)(XW - Y)')$$

$$r_y = diag((X - YW')(X - YW')')$$

Eq. (6) is a standard quadratic programming problem, which can be solved efficiently and easily by many well-established QP solvers, such as the $\mathrm{quadprog}$[2] function in MATLAB.

3.3 Discussion

Since we use an iterative algorithm to solve Eq. (3), we need to initialize variables at first. In this paper, we simply initialize $\mu_i = \delta$ and optimize W first. Since our goal is to find the representative images for cross-domain alignment. At first, no cross-domain alignment function is available, and thus we have to initialize them with the same weight. In fact, besides the self-initialization, we can also consider a warm-start method by using other ZSL approaches to find the representative images and initialize μ_i accordingly.

As introduced above, the complexity of solving Eq. (4) is irrelevant to n_s. However, the QP solver for Eq. (6) has polynomial complexity of n_s. In this case, training a model on a large-scale dataset, like ImageNet is quite challenging. To address this issue, we propose a mini-batch based strategy. For a dataset with n_s training samples, we divide it into M mini batches and each batch has n_s/M samples. We use one batch to solve Eq. (6) to obtain the weights for samples in this batch and then we process each batch in the same way. In this way, the complexity of solving Eq. (6) is reduced from $poly(n_s)$ to $n_s \times poly(b)$ where b the the size of a mini batch which is much smaller than n_s. This strategy turns out to be effective and efficient in practice.

Algorithm 1 Landmark Selection for ZSL

Input: Training set $\{x_i, y_i\}_{i=1}^{n_s}$, parameter λ, α, β, γ, δ;
Output: The compatibility matrix W;
Initialize: $\mu_i = \delta$;
repeat
  Update W by solving Eq. (4);
  for one mini batch in training set **do**
    Update μ_i for this batch by solving Eq. (6);
  end for
until Convergence;
Return W;

At test stage, given a test sample and a set of unseen classes, the compatibility function is defined as:

$$F_{LAST}(x, y; W) = \frac{exp(-\|xW - y\|_2^2)}{\sum_{c=1}^{k_u} exp(-\|xW - y^c\|_2^2)} + \frac{exp(-\|x - yW'\|_2^2)}{\sum_{c=1}^{k_u} exp(-\|x - y^c W'\|_2^2)} \tag{7}$$

which considers the embedding in image feature space and class feature space like the training objective function, and normalizes them for addition by a softmax-like operation.

4 Experiment

4.1 Setting

We use five widely used ZSL benchmark datasets for evaluation. The first dataset is Animals with Attributes2 (AwA2) [Xian *et al.*, 2017] which has 50 animal categories. In AwA2, 40 categories are used as seen classes and the other 10 as unseen classes. The second dataset is aPascal-aYahoo (aPY) [Farhadi *et al.*, 2009]. It has 20 classes from Pascal VOC challenge like "person" and "dog" as the seen classes, and 12 related classes like "centaur" and "wolf" collected from Yahoo search engine. The third dataset is SUN [Patterson and Hays, 2012] scene recognition dataset which has 717 different scenes of which 645 are used as seen classes and the other 72 as unseen classes. The fourth dataset is CUB [Wah *et al.*, 2011] bird fine-grained recognition dataset with 200 kinds of birds of which 150 are used as seen classes and the other 50 as unseen classes. The last dataset is ImageNet [Russakovsky *et al.*, 2015] which is a large-scale dataset with a large number of classes. The widely used $1,000$ classes with about 1.3 images are used as training set. There are another about 20k classes with about 14 million images, which are utilized as the test set. To comprehensively evaluate on ImageNet, we consider different subset of the test set, including classes that are 2-hops (denoted as 2H, $1,509$ classes) and 3-hops (3H, $7,678$ classes) away from the $1,000$ seen classes, the most popular 500 (M500), 1k (M1K), and 5k (M5k) classes, and the least popular 500 (L500), 1k (L1K), and 5k (L5K) classes. For each dataset, some seen class images are used for model

[1] https://uk.mathworks.com/help/matlab/ref/sylvester.html
[2] https://www.mathworks.com/help/optim/ug/quadprog.html

	AwA2		aPY		SUN		CUB		Average	
	ACC	H	ACC	H	ACC	H	ACC	H	ACC	H
DEVISE [Frome *et al.*, 2013]	59.7	27.8	39.8	9.2	56.5	20.9	52.0	32.8	52.00	22.68
CMT [Socher *et al.*, 2013]	37.9	15.9	28.0	19.0	39.9	13.3	34.6	8.7	35.10	14.23
SJE [Akata *et al.*, 2015]	61.9	14.4	32.9	6.9	53.7	19.8	53.9	33.6	50.60	18.68
EZZSL [Romera-Paredes and Torr, 2015]	58.6	11.0	38.3	4.6	54.5	15.8	53.9	21.0	51.33	13.10
ALE [Akata *et al.*, 2016]	62.5	23.9	39.7	8.7	58.1	26.3	54.9	34.4	53.80	23.33
SYNC [Changpinyo *et al.*, 2016]	46.6	18.0	23.9	13.3	56.3	13.4	55.6	19.8	45.60	16.13
LATEM [Xian *et al.*, 2016]	55.8	20.0	35.2	0.2	55.3	19.5	49.3	24.0	48.90	15.93
PSR [Annadani and Biswas, 2018]	63.8	32.3	38.4	21.4	61.4	26.7	56.0	33.9	54.90	28.58
ICINESS [Guo *et al.*, 2018]	64.2	36.3	42.4	23.1	62.9	30.3	59.8	39.4	57.33	32.28
ZKL [Zhang and Koniusz, 2018]	70.5	30.8	45.3	20.5	61.7	25.1	57.1	35.1	58.65	27.88
LAST	**71.0**	**38.4**	**47.2**	**26.7**	**64.7**	**33.2**	**62.8**	**41.9**	**61.43**	**35.05**

Table 1: (Generalized) ZSL performance comparison on benchmarks. ZSL is evaluated by ACC and GZSL is evaluated by H.

	2H	3H	M500	M1k	M5k	L500	L1k	L5k	ALL
CONSE [Norouzi *et al.*, 2013]	7.63	2.18	12.33	8.31	3.22	3.53	2.69	1.05	0.95
CMT [Socher *et al.*, 2013]	2.88	0.67	5.10	3.04	1.04	1.87	1.08	0.33	0.29
LATEM [Xian *et al.*, 2016]	5.45	1.32	10.81	6.63	1.90	4.53	2.74	0.76	0.50
ALE [Akata *et al.*, 2016]	5.38	1.32	10.40	6.77	2.00	4.27	2.85	0.79	0.50
DEVISE [Frome *et al.*, 2013]	5.25	1.29	10.36	6.68	1.94	4.23	2.86	0.78	0.49
SJE [Akata *et al.*, 2015]	5.31	1.33	9.88	6.53	1.99	4.93	2.93	0.78	0.52
ESZSL [Romera-Paredes and Torr, 2015]	6.35	1.51	11.91	7.69	2.34	4.50	3.23	0.94	0.62
SYNC [Changpinyo *et al.*, 2016]	9.26	2.29	15.83	10.75	3.42	5.83	3.52	1.26	0.96
SAE [Kodirov *et al.*, 2017]	4.89	1.26	9.96	6.57	2.09	2.50	2.17	0.72	0.56
LAST	**10.27**	**2.44**	**17.19**	**12.37**	**3.72**	**6.71**	**4.32**	**1.47**	**1.21**

Table 2: ZSL performance comparison on ImageNet.

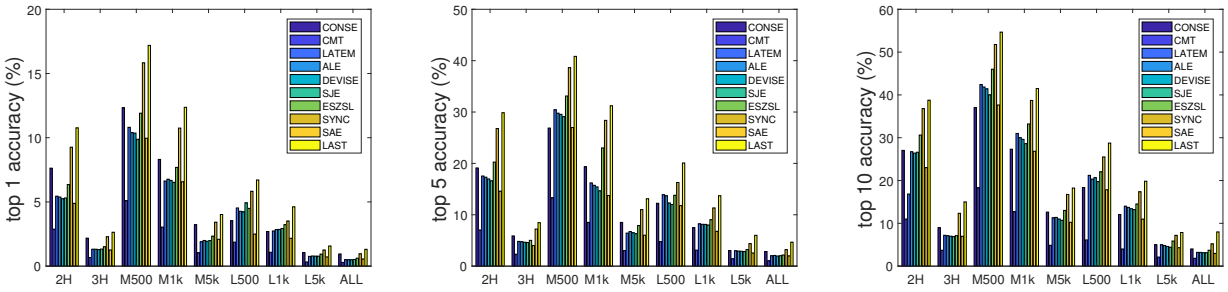


Figure 2: The top-1, top-5, and top-10 accuracy on ImageNet.

training and the other seen class images together with all unseen class images are utilized as the test set. For each image, the ResNet-101 [He *et al.*, 2016] pre-trained on ImageNet is employed as feature extractor producing $2,048$-dimensional image feature. For each class, the class attribute vector is regarded as the label feature vector for the first four datasets. For ImageNet, we use the 500-dimensional word2vec representations for all classes [Changpinyo *et al.*, 2016]. For fair comparison, we make use of the same seen-unseen split, train-test split, image feature, and label feature given by Xian *et al.* [2017]. To evaluation the performance, we consider two tasks. The first task is standard ZSL where a test image comes from only unseen classes. The second task is generalized ZSL where a test image may come from both seen and unseen classes. To evaluate performance, we use the *average per-class top-1 accuracy* and the *harmonic mean* for two tasks respectively. Please refer to [Xian *et al.*, 2017] details.

To implement LAST, we use the following settings. We set the parameters in Eq. (3) as: $\lambda = 1$, $\alpha = \beta = n_s/10$, $\gamma = 0.01$ and $\delta = 0.5$. When using mini-batch based optimization, we set the mini batch size to $1,024$. For both image features and class features, we further perform dimension-wise centralization such that these features have zero mean.

4.2 Result

The comparison between LAST and several state-of-the-art ZSL approaches on four small-scale benchmark datasets is summarized in Table 1, and the comparison on ImageNet is shown in Table 2 and Figure 2. It is clearly observed from the results that LAST outperforms the state-of-the-art ZS-

L approaches with significant and consistent margin, which demonstrates the effectiveness of LAST for ZSL. Besides, we can also obtain the following observations from the results.

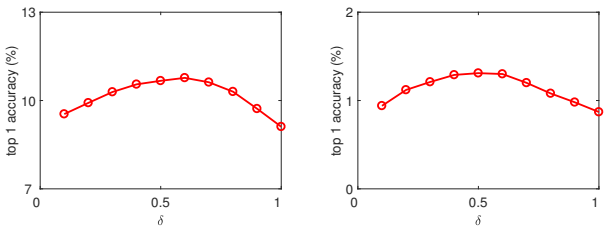
Firstly, among all approaches, LAST is the only one which does not use all images for training. In fact, since each class has only one class feature, it is not reasonable to use unrepresentative images to fit the class feature, which will result in biased model. LAST combines model training and landmark selection such that it is capable of finding the most representative images for model training. In this way, the model can focus on the most important and general properties for aligning heterogeneous spaces, making it perform better for ZSL.

Secondly, not only the conditional alignment which performs alignment for each class, LAST also takes into account the marginal distribution such that it can align heterogeneous feature spaces globally. This property is very important for ZSL. As introduced, ZSL is to build recognition model for previously unseen classes. Therefore, it is necessary that the ZSL model should generalize well. If only conditional distribution is considered, the model may pay too much attention to the details between image feature space and class feature space such that it does not well align two spaces. We notice that ZSL can be regarded as a alignment problem between image feature space and class feature space. Therefore, aligning them globally is essential for building an effective model. One may argue that image feature space and class feature space has zero MMD after feature centralization since W is a simple linear transformation. This is true in a usual case. However, with the landmark selection, it is not guaranteed that the weighted samples still have zero mean. Therefore, it is necessary to combining this term to the objective function.

4.3 Analysis

The parameter δ controls the power of landmark selection. It is interesting to investigate its influence on LAST. Here we use 2H and ALL of ImageNet for analysis. The top 1 accuracy of LAST with respect to different values of δ is shown in Figure 3. When δ is small (say, 0.1), the performance drops because too many images are ignored. In this case, the alignment between heterogeneous spaces is imperfect since too much information is lost. On the other hand, when δ is large (say, 1), the performance is worse than the best performance too. This phenomenon is very important to verify the motivation of LAST. When $\delta = 1$, all images are used with the same weight and no landmark selection is performed. We can observe that any smaller value for δ leads to better performance than the case when $\delta = 1$. This result clearly demonstrates the basic idea of LAST that selecting representative images, instead of using all indiscriminately, is beneficial for ZSL.

In addition, we propose to incorporate a marginal alignment term into LAST by minimizing a MMD loss. To investigate its effectiveness, we compare LAST to the version with only conditional alignment denoted as LAST-C. The comparison one four benchmarks and eight subsets of ImageNet is shown in Figure 4. It can be observe that the marginal alignment significantly and consistently promote the performance, which validates the efficacy of the marginal alignment.



(a) 2H (b) ALL

Figure 3: The effect of δ.

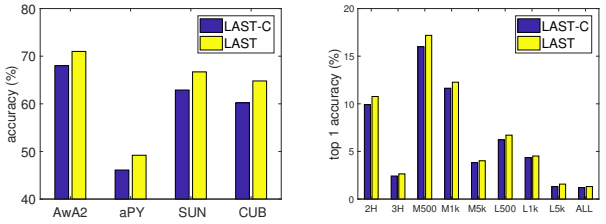


Figure 4: The effect of marginal alignment.

5 Conclusion

In this paper we focus on ZSL. We notice that previous ZSL approaches are mostly based heterogenous feature matching and the matching/compatibility function is constructed from one-vs-all training in which each class feature is combined with all image features for loss computation. However, this operation seems unreasonable since some images may focus on details that the class feature does not reflects. Therefore, the mismatching between them leads to biased ZSL model. To address this issue, we propose a novel approach, termed as LAST. In particular, during model training, LAST simultaneously perform landmark selection which assigns different weights to different samples, aiming at finding the representative cross-domain data instead of using all. In addition, a marginal alignment term is incorporated into LAST to improve its generalization ability. LAST is capable of identifying representative cross-domain features which further lead to better image-class compatibility function. Experiments on several benchmark datasets including ImageNet demonstrate that LAST significantly outperforms the state-of-the-arts.

Acknowledgements

This work was supported by the National Key R&D Program of China (No. 2018YFC0807500), National Natural Science Foundation of China (No. 61571269, 61671196, 61525206), 111 Project (No. D17019), National Natural Science Major Foundation of Research Instrumentation of China (No. 61427808), National Postdoctoral Program for Innovative Talents (No. BX20180172), and the China Postdoctoral Science Foundation (No. 2018M640131).
Email: yuchen.w.guo@gmail.com. Corresponding author: Guiguang Ding, dinggg@tsinghua.edu.cn.

References

[Aggarwal *et al.*, 2014] Charu C. Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and Philip S. Yu. Active learning: A survey. In *Data Classification: Algorithms and Applications*, pages 571–606. 2014.

[Akata *et al.*, 2015] Zeynep Akata, Scott E. Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015.

[Akata *et al.*, 2016] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE TPAMI*, 2016.

[Annadani and Biswas, 2018] Yashas Annadani and Soma Biswas. Preserving semantic relations for zero-shot learning. In *CVPR*, 2018.

[Bartels and Stewart, 1972] Richard H. Bartels and G. W. Stewart. Solution of the matrix equation ax+xb=c [F4] (algorithm 432). *Commun. ACM*, 15(9):820–826, 1972.

[Changpinyo *et al.*, 2016] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016.

[Chen *et al.*, 2018] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *CVPR*, 2018.

[Farhadi *et al.*, 2009] Ali Farhadi, Ian Endres, Derek Hoiem, and David A. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[Felix *et al.*, 2018] Rafael Felix, B. G. Vijay Kumar, Ian D. Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, 2018.

[Frome *et al.*, 2013] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[Fu *et al.*, 2015] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot object recognition by semantic manifold distance. In *CVPR*, 2015.

[Guo *et al.*, 2016] Yuchen Guo, Guiguang Ding, Xiaoming Jin, and Jianmin Wang. Transductive zero-shot recognition via shared model space learning. In *AAAI*, 2016.

[Guo *et al.*, 2017a] Yuchen Guo, Guiguang Ding, Jungong Han, and Yue Gao. Synthesizing samples for zero-shot learning. In *IJCAI*, 2017.

[Guo *et al.*, 2017b] Yuchen Guo, Guiguang Ding, Jungong Han, and Yue Gao. Zero-shot learning with transferred samples. *IEEE TIP*, 26(7):3277–3290, 2017.

[Guo *et al.*, 2018] Yuchen Guo, Guiguang Ding, Jungong Han, Sicheng Zhao, and Bin Wang. Implicit non-linear similarity scoring for recognizing unseen classes. In *IJCAI*, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Huang *et al.*, 2012] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Improving word representations via global context and multiple word prototypes. In *ACL*, 2012.

[Kodirov *et al.*, 2017] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, 2017.

[Lampert *et al.*, 2014] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 2014.

[Norouzi *et al.*, 2013] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *CoRR*, abs/1312.5650, 2013.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, 2010.

[Patterson and Hays, 2012] Genevieve Patterson and James Hays. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012.

[Romera-Paredes and Torr, 2015] Bernardino Romera-Paredes and Philip H. S. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015.

[Russakovsky *et al.*, 2015] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z.g Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[Shrivastava *et al.*, 2016] Abhinav Shrivastava, Abhinav Gupta, and Ross B. Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016.

[Socher *et al.*, 2013] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013.

[Tsai *et al.*, 2016] Yao-Hung Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang. Learning cross-domain landmarks for heterogeneous domain adaptation. In *CVPR*, 2016.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[Wang and Deng, 2018] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 2018.

[Xian *et al.*, 2016] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh N. Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016.

[Xian *et al.*, 2017] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. *CoRR*, abs/1707.00600, 2017.

[Ye and Guo, 2017] Meng Ye and Yuhong Guo. Zero-shot classification with discriminative semantic representation learning. In *CVPR*, 2017.

[Zhang and Koniusz, 2018] Hongguang Zhang and Piotr Koniusz. Zero-shot kernel learning. In *CVPR*, 2018.

[Zhang and Saligrama, 2015] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, 2015.

[Zhang and Saligrama, 2016] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, 2016.