Structure propagation for zero-shot learning

Guangfeng Lin[a,*], Yajun Chen[a], Fan Zhao[a]

[a]*Information science department, Xian University of Technology,*
5 South Jinhua Road, Xi'an, Shaanxi Province 710048, PR China

---

Abstract

The key of zero-shot learning (ZSL) is how to find the information transfer model for bridging the gap between images and semantic information (texts or attributes). Existing ZSL methods usually construct the compatibility function between images and class labels with the consideration of the relevance on the semantic classes (the manifold structure of semantic classes). However, the relationship of image classes (the manifold structure of image classes) is also very important for the compatibility model construction. It is difficult to capture the relationship among image classes due to unseen classes, so that the manifold structure of image classes often is ignored in ZSL. To complement each other between the manifold structure of image classes and that of semantic classes information, we propose structure propagation (SP) for improving the performance of ZSL for classification. SP can jointly consider the manifold structure of image classes and that of semantic classes for approximating to the intrinsic structure of object classes. Moreover, the SP can describe the constrain condition between the compatibility function and these manifold structures for balancing the influence of the structure propagation iteration. The SP solution provides not only unseen class labels but also the relationship of two manifold structures that encode the positive transfer in structure propagation. Experimental results demonstrate that SP can attain the promising results on the AwA, CUB, Dogs and SUN databases.

Keywords: structure propagation, manifold structure, zero-shot learning, transfer learning

---

[*] Corresponding author
Email address: `lgf78103@xaut.edu.cn` (Guangfeng Lin)

1. Introduction

Although deep learning [1] depending on large-scale labeled data training has been widespreadly used for visual recognition, a daunting challenge still exists to recognize visual object ”in the wild”. In fact, in specific applications it is impossible to collect all class data for training deep model, so training and testing class sets are often disjoint. The main idea of ZSL is to handle this problem by exploit the transfer model via the redundant relevance of the semantic description. Furthermore, in ZSL, testing class images can be mapped into the semantic or label space by transfer model for recognizing objects of unseen classes, from which samples are not available or can not be collected in training sets. Many ZSL methods [2] [3] [4] [5] [6] [7] [8] attempt to model the interaction relationship on the cross-domain (e.g. text domain or image domain) via transfer model to classifying objects of unseen classes by the aid of the semantic description of unseen classes and seen classes, from which labeled samples can be used. For example, ’pig’ is a unseen class in testing image sets, while ’zebra’ is a seen class in training image sets. These classes both have the related semantic description(e.g. attribute is ’has tail’). Therefore, ZSL can construct a knowledge transfer model between ’zebra’ and ’has tail’ in training sets, and then, ’pig’ can be mapped into the semantic or label space by this model for recognizing ’pig’ in testing image sets.

To recognize unseen classes from seen classes, ZSL needs face to two challenges [7]. One is how to utilize the semantic information for constructing the relationship between unseen classes and seen classes, and other is how to find the compatibility among all kinds of information for obtaining the optimal discriminative characteristics on unseen classes.

To handle the first challenge, visual attributes [9] [10] [11] and text representations [12] [13] [14] have been used for linking unseen and seen classes. These semantic information can not only be regarded as a middle bridge for associating visual images and class labels [14] [15] [16] [17] [18] [19] [20], but also be transformed into new representations corresponding to the more suitable relation between unseen and seen classes by Canonical Correlation Analysis (CCA)[21] or Sparse Coding (SC)[22] [23]. To address the second challenge, the classical method as baseline is the probability model for visual attributes predicting unseen class labels [16]. For implementing the discriminative classification, the recent methods have two tendencies. some methods have built the linear [12] [24] [25], nonlinear [14] [26] or hybrid [19] [23] compatibility function between image domain and semantic domain (for example,text domain), others have further considered the manifold structure of semantic classes for constraining the compatibility function [7]. However, the manifold structure of image classes that can enhance the connection between unseen classes and seen classes is often neglected, because unseen classes make the manifold structure of image classes to be uncertain.

In this paper, our motivation is inspired by structure fusion [27] [28] [29] [30] [31] [32] [33]for jointly dealing with two challenges. The intrinsic manifold structure is crucial for object classification. However, in fact, we only can attain the observation data of the manifold structure, which can represent different aspects of the intrinsic manifold structure. For recovering or approximating the intrinsic structure, we can fuse various manifold structures from observation data. Based on the above idea, we try to capture different manifold structures in image and semantic space for improving the recognition performance of unseen classes in ZSL. We view the weighted graph of object classes in semantic or image space as the different manifold structure. In the weighted graph of semantic space (the manifold structure of semantic classes), nodes are corresponding to semantic representations (e.g. attributes[9], word vectors [34], GloVe [35] or Hierarchical embeddings [24]) of object classes and these weights of edges describe the distance or similarity relationship of nodes. In the weighted graph of image space(the manifold structure of image classes), it is difficulty to obtain some certain nodes and weights because we do not know labels of unseen classes. Therefore, we expect to construct the compatibility function for predicting labels of unseen classes by building the manifold structure of image classes. On the other end, we attempt to find the relevance between the manifold structure of semantic classes and that of image classes in model space for encoding the influence between the negative and positive transfer, and further make the better compatibility function for classifying unseen class objects. Finally, we iterate the above process to converge the stable state for obtaining the discriminative performance, and in this iteration process manifold structure can be propagated on unseen classes. Model space corresponding to visual appearances is the jointed projection space of semantic space and image space, and can preserve the respective manifold structure. Therefore, we respectively define phantom object classes (the coordinates of classes in the model space are optimized to achieve the best performance of the resulting model for the real object classes in discriminative tasks [7].) and real object classes corresponding to all classes in model space. Figure 1 illustrates the idea of the proposed method conceptually.

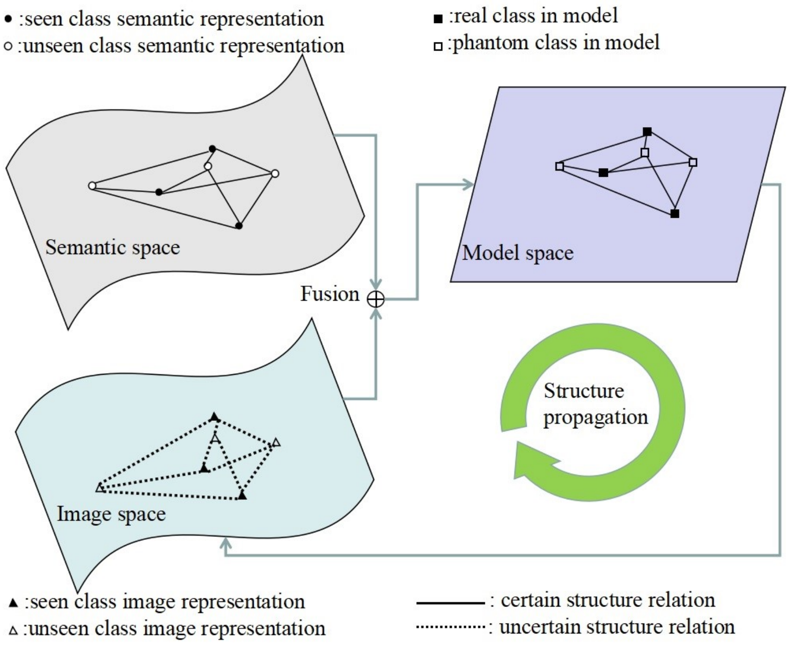
● :seen class semantic representation
○ :unseen class semantic representation
■ :real class in model
□ :phantom class in model
Semantic space
Model space
Fusion
Structure propagation
Image space
▲ :seen class image representation
△ :unseen class image representation
: certain structure relation
: uncertain structure relation

Figure 1: The illustration of structure propagation for zero-shot learning.

In our main contribution, a novel idea is to recover or approximate the intrinsic manifold structure from seen classes to unseen classes by fusing the different space manifold structure for handling the challenging unseen classes recognition. Specifically, we demonstrate how to construct the projected manifold structure for real and phantom class in model space, and how to constrain the compatibility function and the relationship of the manifold structure for the positive structure propagation. In the experiment, we evaluate SP on four benchmark datasets for ZSL. Experimental results are promising for improving the recognition performance of unseen class objects.

2. Related Works

ZSL can bridge the gap among the different domains to recognize unseen class objects by semantic information, which includes the class label and usually can be called the semantic embedding of class labels. These semantic embeddings can come from vision (attributes [10]) and language information (text [14]) by the manual annotation [36], machine learning [37]or data mining [38]. In term of the transformation relationship of different embedding, recent ZSL methods mainly fall into linear embedding, nonlinear embedding and similarity embedding.

Linear embedding implements the linear transformation method among different embedding spaces for learning the relevance between unseen class objects and class labels. In classical methods, the first step maps image feature to semantic space, and then the second step recognizes image object by class labels in the semantic space [36] [24]. In recent methods,the above steps are combined into a unified framework. Especially, some representation methods, which are max-margin learning learning label embedding [39], ranking objective[12], weighted approximate ranking objective [40], full weight to the top of the ranked list [24]and risk minimization formulation with regularization term [20], can obtain the compatibility of latent space by addressing image and class bi-linear embedding for attaining the better recognition performance of unseen class objects.

Nonlinear embedding can realize the nonlinear mapping of the embedding space for building the compatibility function or classifier. There mainly are three ways for constructing the nonlinear mapping. The first way is a piece-wise linear compatibility for modeling the different characteristic of the embedding [26], and is convenience for computing the transformation matrix of the nonlinear compatibility function in the cross-domain. The second way is a nonlinear hyperbolic tangent activation for learning from image to semantic space of words [14] or computing inner product of hypothetical space [41], and is suitable for the threshold transformation in the cross-domain. The third way is a kernel function between two images for defining the discriminant function of the intra-modal label transfer [41], and is fit for space metric in the same domain.

Different from the above embeddings, similarity embedding builds the classifier by the similarity metrics, which mostly include structure learning or class-wise similarities. By structure learning, similarity embedding learns a joint latent space in cross domain for fitting each sample by dictionary learning and improving the recognition performance by bilinear classifier [42] [23]. Via class-wise similarities, unseen classes associate with seen classes for enhancing the unseen object classification and the domain compatibility [43] [44] [7]. Recently, dual visual-semantic mapping paths[45] can capture and refine the semantic space manifold structure (it can be described by the similarity metric) to enhance the transfer ability of visual-semantic mapping for unseen classes classification. In our approach, the similarity metric is extended from semantic space to image space, we attempt to find the relationship of similarities (manifold structure in the different space) for constraining the compatibility function, and further capture to the positive structure propagation for the significantly improvement of the unseen object classification. Figure 2 shows a flowchart for describing the different steps of the proposed method.

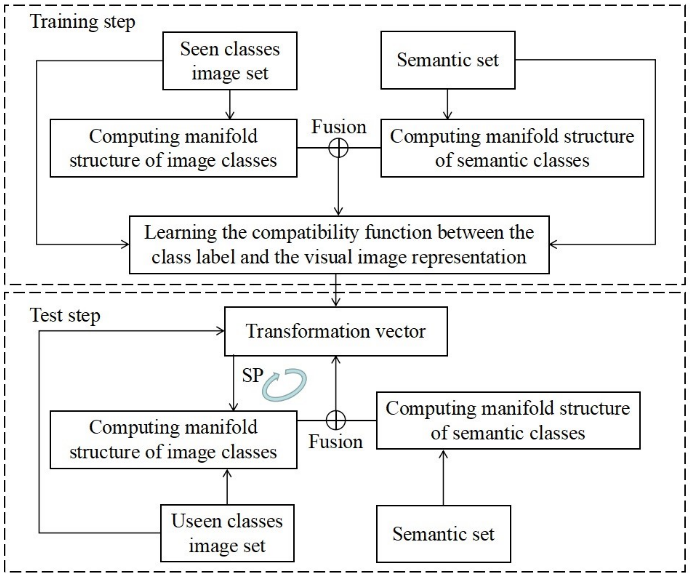
Training step
Seen classes image set
Semantic set
Computing manifold structure of image classes
Fusion
Computing manifold structure of semantic classes
Learning the compatibility function between the class label and the visual image representation
Test step
Transformation vector
SP
Computing manifold structure of image classes
Fusion
Computing manifold structure of semantic classes
Useen classes image set
Semantic set

Figure 2: The flowchart of structure propagation (SP) for zero-shot learning (ZSL).

3. Structure propagation

In ZSL, we have training data set $\mathscr{D} = \{(x_n \in R^D, y_n)\}_{n=1}^N$, in which x_n is image representation (it can be extracted based on deep model, and the detail is described in section 4.2.) and $y_n (n = 1, ..., N)$ is the class label in the seen class set $\mathscr{S} = \{s|s = 1, ..., S\}$. We can denote the unseen class set as $\mathscr{U} = \{u|u = S+1, ..., S+U\}$. $a_c \in R_D$ is the linear transformation vector of the $c \in \{\mathscr{S} \bigcup \mathscr{U}\}$ class. Table 1 lists the important notations used in this paper.

3.1. Classification model and manifold structure

We construct a pair-wise linear classifier [7]in the visual image feature space, and determinate a estimated label $\hat{y}$ to a feature x by the following formula.

$$\hat{y} = \arg\max_c a_c^T x, \tag{1}$$

here, $a_c \in R^D$ is not only the transformation vector of the feature x, but also the representation of the class c in model. In other words, the above formula can describe the pair-wise linear relation between the feature space and the class label space for characterizing the class representation in the model.

To measure the manifold structure, we can compute the similarity of the related representation in the homogeneous space, which has the same scale and metric. To this end we respectively build a bipartite graph between unseen classes and seen classes in semantic space and image space (this space includes all image representations). In these bipartite graphes, nodes are corresponding to unseen classes or seen classes, and weights of these nodes connect unseen classes with seen classes. Because we focus on the transfer relation between unseen classes and seen classes, no connection exists in unseen classes or seen classes. Supposing $G_b < V_b, E_b >$ can denote the manifold structure of semantic classes. Here, $V_b = V_{bs} \bigcup V_{bu}$ and $\emptyset = V_{bs} \bigcap V_{bu}$. E_b includes connections between V_{bs}(seen classes set in semantic space) and V_{bu}(unseen classes set in semantic space). Therefore, similarity is regarded as the weight between nodes, which can be defined as following.

$$w_{su}^{(b)} = \frac{\exp(-d(b_s, b_u))}{\sum_{u=1}^{U} \exp(-d(b_s, b_u))}, \tag{2}$$

Table 1: List of mathematical notations.

Notation	**Description**
$\mathscr{D} = \{(x_n \in R^D, y_n)\}_{n=1}^N$	Training data set includes the image representation x_n and the class label y_n in nth sample
$\mathscr{S} = \{s \| s = 1, ..., S\}$	The class label space of seen classes
$\mathscr{U} = \{u \| u = S+1, ..., S+U\}$	The class label space of unseen classes
$c \in \{\mathscr{S} \bigcup \mathscr{U}\}$	The label of any class in $\mathscr{S} \bigcup \mathscr{U}$
a_c	The transformation vector of the linear model or any real class c representation in model space
v_s	The phantom class representation corresponding to the seen class s in model space
v_u	The phantom class representation corresponding to the unseen class u in model space
b_s	The semantic representation of the seen class s
b_u	The semantic representation of the unseen class u
x_s	The image representation of the seen class s
x_u	The image representation of the unseen class u
$w_{su}^{(b)}$	Similarity between the seen class s and the unseen class u in semantic representation b space
$w_{su}^{(x)}$	Similarity between the seen class s and the unseen class u in semantic representation x space

here, b_s is the semantic representation (is the vector feature from different semantic sources, and the detail is described in section 4.2.) of the seen class s, and b_u is the semantic representation of the unseen class u. $w_{su}^{(b)}$ is the weight (the similarity) between the seen class s and the unseen class u in semantic representation b space. $d(b_s, b_u)$ is a distance metric [7], and can be defined as following.

$$d(b_s, b_u) = (b_s - b_u)^T \Sigma_b^{-1} (b_s - b_u), \tag{3}$$

here, $\Sigma_b = \sigma_b I$ can be learned from the semantic representation by cross-validation (We alternately divide the training classes set into two part in according with the proportion between the training classes set and the test classes set. One part is to learn the model, and anther is to validate the model. We give the range of σ_b, which is form 2^{-5} to 2^5, and select the parameter corresponding to the best result as the value of σ_b.)

Like the manifold structure of semantic classes, we build a bipartite graphes $G_x < V_x, E_x >$ for the manifold structure of image classes. Here, $V_x = V_{xs} \bigcup V_{xu}$ and $\emptyset = V_{xs} \bigcap V_{xu}$. E_x includes the connections between V_{xs}(seen classes set in image space) and V_{xu}(unseen classes set in image space). In term of (2) and (3), we can define the weight on G_x as following.

$$w_{su}^{(x)} = \frac{\exp(-d(x_s, x_u))}{\sum_{u=1}^{U} \exp(-d(x_s, x_u))}, \tag{4}$$

$$d(x_s, x_u) = (x_s - x_u)^T \Sigma_x^{-1} (x_s - x_u), \tag{5}$$

here, $\Sigma_x = \sigma_x I$ can be learned from the image representation by cross-validation (It is the same procedure like σ_b learning.). In image space, the differentiation compared with the semantic space is that x_u is not determined because of unseen classes, while x_s can be obtained from training data by computing the mean value of the seen class. The way to produce the center of the class as a representation is simple for convenient computation, and it is reasonable to preserve the base characteristic of image representation according with the distribution of the same class. x_u can be attained by pre-classification of unseen classes (the detail in the next section).

In (1), a_c is the transformation vector, and also is the class representation in model space. In (2), b_s and b_u is the class representation in semantic space. In (4), x_s and x_u is the class representation in image space. We expect to construct the link among these space by v_s and v_u, which are respectively the phantom class of seen or unseen classes in model. For preserving the manifold structure of two bipartite graphes and aligning the image, the semantic and the model space, we build the optimization formula under the condition of the distortion error minimization, which is defined as following.

$$\begin{aligned}(a_c, v_u, \vec{\beta}) = \arg \min_{a_c, v_u, \vec{\beta}} \| a_c - \sum_{u=1}^{U} \vec{\beta}^T \begin{bmatrix} w_{su}^{(x)} & w_{su}^{(b)} \end{bmatrix}^T v_u \\ - \sum_{s=1}^{S} \vec{\gamma}^T \begin{bmatrix} w_{ss}^{(x)} & w_{ss}^{(b)} \end{bmatrix}^T v_s \|_2^2, \\ s.t. \quad \vec{\beta}^T \vec{\mathbf{1}} = 1, \vec{\gamma}^T \vec{\mathbf{1}} = 1, 0 \leq \beta_i \leq 1, 0 \leq \gamma_i \leq 1 \quad (i = 1, 2)\end{aligned} \tag{6}$$

here,$\vec{\beta} = \begin{bmatrix} \beta_1 & \beta_2 \end{bmatrix}^T$, $\vec{\gamma} = \begin{bmatrix} \gamma_1 & \gamma_2 \end{bmatrix}^T$,and$\vec{\mathbf{1}} = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$. Because no connection exists between unseen classes or seen classes in tow bipartite graphes, $w_{ss}^{(b)} = 0$ and $w_{ss}^{(x)} = 0$. Therefore, (6) can be transformed into the following formula.

$$\begin{aligned}(a_c, v_u, \vec{\beta}) = &\arg \min_{a_c, v_u, \vec{\beta}} \| a_c - \sum_{u=1}^{U} \vec{\beta}^T \begin{bmatrix} w_{su}^{(x)} & w_{su}^{(b)} \end{bmatrix}^T v_u \|_2^2, \\ s.t. \quad &\vec{\beta}^T \vec{\mathbf{1}} = 1, 0 \leq \beta_i \leq 1 \quad (i = 1, 2)\end{aligned} \tag{7}$$

The analytical solution of (7) can find the relation between a_c and v_u.

$$\begin{aligned}a_c = &\sum_{u=1}^{U} \vec{\beta}^T \begin{bmatrix} w_{su}^{(x)} & w_{su}^{(b)} \end{bmatrix}^T v_u, \\ s.t. \quad &\vec{\beta}^T \vec{\mathbf{1}} = 1, 0 \leq \beta_i \leq 1 \quad (i = 1, 2)\end{aligned} \tag{8}$$

here, $\forall c \in \{1, 2, ..., S + U\}$.

3.2. Phantom classes and structure relation learning

For obtaining phantom class $v_u (u = 1, ..., U)$ and the manifold structure of the weight coefficient vector β, we further reformulate the optimization formula for one-versus-other classifier [7].

$$\begin{aligned}(v_1, ..., v_U, \vec{\beta}) = \arg &\min_{v_1, ..., v_U, \vec{\beta}} \sum_{c=1}^{S} \sum_{n=1}^{N} \ell(x_n, \mathbb{I}_{y_n, c}, a_c) + \frac{\lambda}{2} \sum_{c=1}^{S} \| a_c \|_2^2, \\ s.t. \quad a_c = &\sum_{u=1}^{U} \vec{\beta}^T \begin{bmatrix} w_{su}^{(x)} & w_{su}^{(b)} \end{bmatrix}^T v_u, \vec{\beta}^T \vec{\mathbf{1}} = 1, 0 \leq \beta_i \leq 1 \quad (i = 1, 2)\end{aligned} \tag{9}$$

here, the first term of formula (9) is the squared hinge loss, which can be defined as $\ell(x_n, \mathbb{I}_{y_n,c}, a_c) = \max(0, 1 - \mathbb{I}_{y_n,c} a_c x_n)$. $\mathbb{I}_{y_n,c} \in \{-1, 1\}$ determinates whether or not $y_n = c$. The second term of formula (9) is a_c of a regularization tern, which avoids over-fitting problem on the pair-wise linear classifier for modeling the relationship between the class label and the image representation. However, in formula (9), we can not determinate the value of $w_{su}^{(x)}$, which can be computed by the certain class label, because the class label of unseen classes can not be got in image space. Therefore, we set the initial value of $w_{su}^{(x)}$ to 0. In other words, we do not consider the manifold structure of the image class in the initial state, and then can complete the optimization of formula (9) by solving the quadratic programming problem. When we can categorize the unseen class, and obtain the image class representation by computing the mean value of the same class image representation, in the next iteration computation the updated $w_{su}^{(x)}$ can be considered into the optimization of formula (9). Although the manifold structure of the image class can be successfully leaded into formula (9), its influence could have two aspects for recognizing the unseen class. One is the positive effect, which is the mostly correct classification of unseen classes because of the positive structure propagation in each iteration. The other is the negative role, which is the mainly uncorrect classification of unseen classes due to the negative structure propagation. The former situation is our expectation. Therefore, we reformulate formula (9) by added the third term as following.

$$\begin{aligned}(v_1, ..., v_U, \vec{\beta}) = \arg & \min_{v_1,...,v_U,\vec{\beta}} \sum_{c=1}^{S} \sum_{n=1}^{N} \ell(x_n, \mathbb{I}_{y_n,c}, a_c) + \frac{\lambda}{2} \sum_{c=1}^{S} \|a_c\|_2^2 \\ & + \frac{\gamma}{2} \|\beta_1 W^x - \beta_2 W^b\|_2^2, \\ & s.t. \quad a_c = \sum_{u=1}^{U} \vec{\beta}^T \begin{bmatrix} w_{su}^{(x)} & w_{su}^{(b)} \end{bmatrix}^T v_u, \\ & \vec{\beta}^T \vec{\mathbf{1}} = 1, 0 \leq \beta_i \leq 1 \quad (i = 1, 2)\end{aligned} \tag{10}$$

here, w_{su}^x is the element of the matrix W^x, and w_{su}^b is the element of the matrix W^b. The third term of formula (10) is the constraint of the manifold structure similarity for preventing the negative structure propagation in image space. The alternating optimization can be implemented for minimizing the formula (10) with respect to $\{v_u\}_{u=1}^U$

and $\vec{\beta}$ by solving the quadratic programming problem.

To depict the whole process of the structure propagation mechanism, we show the pseudo code of the proposed SP algorithm in Algorithm 1, which includes three steps. The first step (line 1) computes the similarity matrix to represent the manifold structure of semantic classes. The second step (line 2) initializes the similarity matrix related to the manifold structure of image classes. The third step includes phantom classes $\{v_u\}_{u=1}^{U}$, weight coefficients $\vec{\beta}$ and the classification of unseen object classes updating by each iteration (from line 3 to line 9). Structure propagation can be completed by the whole iteration computation.

Algorithm 1 The pseudo code of the SP algorithm

```
Input: $\mathscr{D} = \{(x_n \in R^D, y_n)\}_{n=1}^{N}$,$b_s$ and $b_u$ (input data)
Output: $y_P^*$ ($P$ is the total iteration number )
1: Computes the similarity matrix $W_{(b)}$ on the semantic representation by (2)
2: Setting the similarity matrix $W_{(x)}$ to zero matrix on the image representation
3: for $1 < t < P$ do
4:     Solving $\{v_u\}_{u=1}^{U}$ and $\vec{\beta}$ by alternately optimizing (10)
5:     Computing $a_c$ according to (8)
6:     Computing $\hat{y}$ by (1) and obtaining the class label $y_t^*$ of the unseen class corresponding to the semantic class
7:     Computing the mean value of each image class as the image class representation $x_s$ and $x_u$
8:     Computing and updating the similarity matrix $W_{(x)}$ on the image representation by (4)
9: end for
```

3.3. Multi-semantic structure fusion

To address multi-semantic structure fusion, we produce w_{su}^{b} by the linear fusion of multi-semantic structure as following.

$$w_{su}^{b} = \vec{\eta}^{T} \begin{bmatrix} w_{su}^{(ba)} & w_{su}^{(bw)} & w_{su}^{(bg)} & w_{su}^{(bh)} \end{bmatrix}^{T}, \tag{11}$$

here, $\vec{\eta} = \begin{bmatrix}\beta_2 & \beta_3 & \beta_4 & \beta_5\end{bmatrix}^T$. $w_{su}^{(ba)}$, $w_{su}^{(bw)}$, $w_{su}^{(bg)}$, and $w_{su}^{(bh)}$ are respectively corresponding to attributes (att)[9], word vectors(w2v) [34], GloVe (glo)[35] and Hierarchical embeddings (hie)[24]. We can bring formula (11) into formula (10) for handling the multi-semantic structure fusion as following.

$$\begin{aligned}(v_1, ..., v_U, \vec{\beta}) = &\arg \min_{v_1,...,v_U,\vec{\beta}} \sum_{c=1}^{S}\sum_{n=1}^{N} \ell(x_n, \mathbb{I}_{y_n,c}, a_c) + \frac{\lambda}{2}\sum_{c=1}^{S} \|a_c\|_2^2 \\ &+ \frac{\gamma}{2}\|\beta_1 W^x - \beta_2 W^{ba} - \beta_3 W^{bw} - \beta_4 W^{bg} - \beta_5 W^{bh}\|_2^2, \\ s.t. \quad a_c = &\sum_{u=1}^{U} \vec{\beta}^T \begin{bmatrix} w_{su}^{(x)} & w_{su}^{(ba)} & w_{su}^{(bw)} & w_{su}^{(bg)} & w_{su}^{(bh)} \end{bmatrix}^T v_u, \\ &\vec{\beta}^T \vec{\mathbf{1}} = 1, 0 \le \beta_i \le 1 \quad (i = 1, ..., 5)\end{aligned} \tag{12}$$

here, $\vec{\beta} = \begin{bmatrix}\beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5\end{bmatrix}^T$,$w_{su}^{(ba)}$, $w_{su}^{(bw)}$, $w_{su}^{(bg)}$, and $w_{su}^{(bh)}$ are respectively the element of $W_{su}^{(ba)}$, $W_{su}^{(bw)}$, $W_{su}^{(bg)}$, and $W_{su}^{(bh)}$. If there are the more semantic information, we can consider these semantic information for ZSL by the similar way of formula (12). Like Algorithm 1, (12) has the similar optimization solving process for considering multi-semantic information in ZSL.

3.4. Complexity analysis

Formula (10) can be solved by alternately quadratic programming, which of the complexity includes two parts. In the first part, when $\vec{\beta}$ is fixed, formula (10) is related to $\{v_u\}_{u=1}^{U}$ of a quadratic programming problem, which of the complexity is $O(U^3)$ for the worst [46]. In the second part, while $\{v_u\}_{u=1}^{U}$ is fixed, formula (10) is corresponding to $\vec{\beta}$ of a quadratic programming problem, which of the complexity is $O(k^3)$ (k is the dimension of $\vec{\beta}$)for the worst [46]. Given the proposed algorithm SP needs P iterations, it's complexity is $O(PU^3 + Pk^3)$.

4. Experiment

4.1. Datasets

For evaluating the proposed algorithm SP, we carry out the experiment in four challenging datasets, which are Animals with Attributes (AwA)[16], CUB-200-2011 Birds

(CUB)[47], Stanford Dogs (Dogs)[48], and SUN Attribute (SUN)[49]. These datasets can be used for fine-grained recognition (CUB and Dogs) or non-fine-grained recognition (AwA and SUN) in ZSL. In semantic space, AwA and CUB respectively are described by att[9],w2v [34],glo[35] and hie[24], while Dogs is represented by w2v [34],glo[35] and hie[24]. SUN is only depicted by att[9]. Figure 3 shows image examples in CUB-200-2011 Birds database. Tab.2 provides the statistics and the extracted features for these datasets.

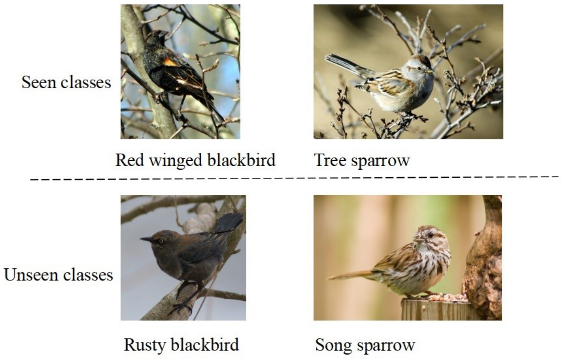

Figure 3: Image examples in CUB-200-2011 Birds database.

4.2. Image and semantic feature

In our SP [1] method, image feature and semantic feature are the main support for modeling ZSL. Deep learning feature can learn the discriminative characteristic of objects based on large scale database. In addition, for conveniently comparing with the state-of-art methods, we adopt image feature provided by [24]. In one word, image feature is the outputs (1024 dimension feature vector) of the pre-trained GoogleNet[50], which can process the whole image as inputs. These images are not pre-processed in any way. In the semantic space, there are four ways to extract the related feature.

---

[1] source code:https://github.com/lgf78103/Structure-propagation-for-zero-shot-learning.

Table 2: Datasets statistics and the extracted feature in experiments.

Datasets	**Number of seen classes**	**Number of unseen classes**	**Total number of images**	**Semantic feature /dimension**	**Image feature /dimension**
AwA	40	10	30473	att/85, w2v/400, glo/400, hie/about 200.	Deep feature based on GoogleNet[50] /1024
CUB	150	50	11786	att/312, w2v/400, glo/400, hie/about 200.	Deep feature based on GoogleNet[50] /1024
Dogs	85	28	19499	N/A, w2v/400, glo/400, hie/about 200.	Deep feature based on GoogleNet[50] /1024
SUN	645	72	14340	att/102, N/A, N/A, N/A.	Deep featurebased on GoogleNet[50] /1024

The first way is the distinguishing vector feature of objects (att) from attributes [9] by human annotation and judgment, which have been completed for collecting data on AwA, CUB, and SUN except Dogs. The second ways is word vectors(w2v) based on a two-layer neural network to predict words through a text document[34]. The third way is GloVe (glo) based on co-occurrence statistics of words from a large unlabel text corpora [35]. The forth way is hierarchical embeddings (hie) based on vectorial class structure from the class hierarchical relationship such as WordNet[24][51]. The w2v, glo, and hie are also provided by [26] to facilitate contrast to the state-of-art methods. In addition, we can extend the other types of visual features in the proposed method, as will be studied for feature fusion in future work.

4.3. Comparison methods

In this paper, there are three methods as the baseline for comparing with the proposed SP method because of the semantic structure mining. The first method is structured joint embedding (SJE) [24], which can build the bilinear compatibility function with the consideration of the structured output space for predicting the label of the unseen class. The second method is latent embedding model (LatEm)[26],which can construct the pair-wise bilinear (nonlinear) compatibility function according to model number selection for recognizing unseen classes. The third method is synthesized classifiers (SynC)[7], which can make nonlinear compatibility function with manifold structure in semantic space for combining the base classifier in ZSL.

4.4. Classification and validation protocols

Classification accuracy is average value of all test class accuracy in each database. Because the learned model involves four parameters, which are λ, γ , σ_b and σ_x (respectively are in formula (3) and (5)) in formula (10). We alternately divide the training classes set into two part in according with the proportion between the training classes set and the test classes set. One part is to learn the model, and anther is to validate the model. Firstly, we set σ_b and σ_x to 1, and obtain γ and λ corresponding to the best result in γ (form 2^{-24} to 2^{-9}) and λ (form 2^{-24} to 2^{-9}) by cross validation. Secondly, we learn σ_b and σ_x corresponding to the best result in σ_b and σ_x (form 2^{-5} to 2^{5}) by cross validation.

4.5. AwA

Animals with Attributes (AwA)[16] is popularly used for ZSL. The characteristic of this dataset is the large scale and the small categories. We extract the deep feature of the image based on the pre-trained GoogleNet[50], the continuous vector (att) for clustering attributes in semantic classes [9], the word vector (w2v)for describing words in the specific context[34], the glover vector (glo) for gathering the co-occurrence words statistics in the given document[35], and the hierarchy vector (hie) for measuring the distance of the hierarchy structure in semantic classes [24]. We have two configurations for ZSL in AwA. One is the comparison SP with the state-of-art methods in each semantic space, the other is the comparison experiment of the fusion methods in multi-semantic space. Figure 4 indicates the experimental comparison of the different method in the various semantic space of AwA. Table 3 shows the performance of the structure propagation (the proposed SP method) greatly outperforms that of other three methods. The highest and the lowest improvement of SP are respectively 26.2% and 10.9% to compare with SJE, 16.3% and 4.6% to contrast to LatEm, or 24.5% and 10.1% to contradistinguish SynC. Table 4 demonstrates the performance of the SP fusion method is better than that of other three fusion methods. Specifically, when the fusion includes the supervised (att) and unsupervised (w2v, glo,and hie) semantic representation, the accuracy of SP can be increased by 11.5% than SJE, 9.3% than LatEm, and 7.4% than SynC. While the fusion only contains the unsupervised semantic description (w2v, glo,and hie), the precision of SP can be enhanced by 21.3% than SJE, 15.2% than LatEm, and 12.3% than SynC.

Table 3: Comparison of SP method with SJE[24], LatEm[26] and SynC[7] in each semantic space, average per-class Top-1 accuracy (%)of unseen classes is reported based on the same data configurations, same images and semantic features in AwA.

Semantic feature	**SJE**	**LatEm**	**SynC**	**SP**
att	66.7	71.9	69.3	**84.3**
w2v	51.2	61.1	52.9	**77.4**
glo	58.8	62.9	53.4	**70.5**
hie	51.2	57.5	52.0	**62.1**

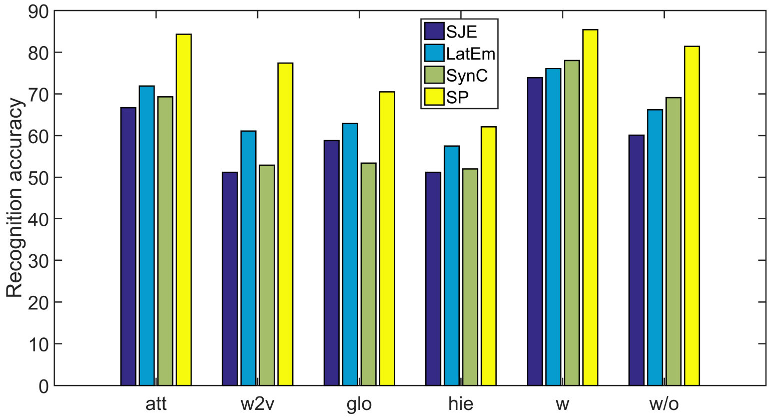

Figure 4: Comparison of SP method with SJE[24], LatEm[26] and SynC[7] in att, w2v, glo, hie, w (the fusion includes att, w2v, glo and hie) and w/o (the fusion contains w2v, glo and hie), average per-class Top-1 accuracy of unseen classes is reported based on the same data configurations, same images and semantic features in AwA.

4.6. CUB

CUB-200-2011 Birds (CUB)[47] is generally utilized for the fine-grained recognition. The scale of this dataset is smaller than AwA, but there are the more categories in CUB. Like AwA, we extracted the deep feature of images in CUB, and we get att, w2v, glo and hie by the semantic description of CUB. Two configurations are respectively the non-fusion and fusion methods comparison in the single and multi-semantic space. Figure 5 indicates the experimental comparison of the different method in the various semantic space of CUB. Table 5 shows that the performance of the structure propagation (the proposed SP method) outperforms that of other three methods. The highest and the lowest improvement of SP are respectively 9.1% and 1.7% to compare with SJE, 6.3% and 0.1% to contrast to LatEm, or 4.3% and 0.2% to contradistinguish SynC. Table 6 demonstrates that the performance of the SP fusion method is

Table 4: Comparison of SP method with SJE[24], LatEm[26] and SynC[7] for multi-semantic fusion, average per-class Top-1 accuracy (%)of unseen classes is reported based on the same data configurations, same images and semantic features in AwA. w: the fusion includes att, w2v, glo and hie, while w/o: the fusion contains w2v, glo and hie.

Fusion	**SJE**	**LatEm**	**SynC**	**SP**
w	73.9	76.1	78.0	**85.4**
w/o	60.1	66.2	69.1	**81.4**

slightly better than that of other three fusion methods. Specifically, when the fusion includes the supervised (att) and unsupervised (w2v, glo,and hie) semantic representation, the accuracy of SP can be increased by 2.4% than SJE, 6.7% than LatEm, and 5.3% than SynC. While the fusion only contains the unsupervised semantic description (w2v, glo,and hie), the precision of SP can be enhanced by 5.4% than SJE, 0.4% than LatEm, and 0.1% than SynC.

Table 5: Comparison of SP method with SJE[24], LatEm[26] and SynC[7] in each semantic space, average per-class Top-1 accuracy(%) of unseen classes is reported based on the same data configurations, same images and semantic features in CUB.

Semantic feature	**SJE**	**LatEm**	**SynC**	**SP**
att	50.1	45.5	47.5	**51.8**
w2v	28.4	31.8	32.3	**32.5**
glo	24.2	32.5	32.8	**33.3**
hie	20.6	24.2	22.7	**24.3**

4.7. Dogs

Stanford Dogs (Dogs)[48] is also usually a benchmark dataset for fine-grained recognition. Dogs is middle between AwA and CUB about the scale and the cate-

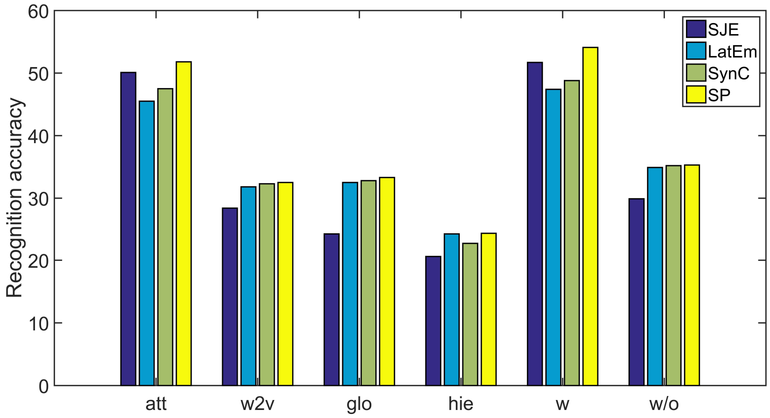

Figure 5: Comparison of SP method with SJE[24], LatEm[26] and SynC[7] in att, w2v, glo, hie, w (the fusion includes att, w2v, glo and hie) and w/o (the fusion contains w2v, glo and hie), average per-class Top-1 accuracy of unseen classes is reported based on the same data configurations, same images and semantic features in CUB.

gory number. We use the same method for extracting the deep feature of images and semantic class features (w2v, glo and hie). We also carry out the experiment in non-fusion methods in the single semantic space and fusion methods in the multi-semantic space. Figure 6 indicates the experimental comparison of the different method in the various semantic space of Dogs. Table 7 shows the performance of the structure propagation (the proposed SP method) outperforms that of other three methods. The highest and the lowest improvement of SP are respectively 15.6% and 8.1% to compare with SJE, 12.5% and 7.2% to contrast to LatEm, or 11.5% and 1.3% to contradistinguish SynC. Table 8 demonstrates the performance of the SP fusion method is obviously better than that of other three fusion methods. Specifically, While the fusion only contains the unsupervised semantic description (w2v, glo,and hie), the precision of SP can be enhanced by 13% than SJE, 11.8% than LatEm, and 11.8% than SynC.

Table 6: Comparison of SP method with SJE[24], LatEm[26] and SynC[7] for multi-semantic fusion, average per-class Top-1 accuracy(%) of unseen classes is reported based on the same data configurations, same images and semantic features in CUB. w: the fusion includes att, w2v, glo and hie, while w/o: the fusion contains w2v, glo and hie.

Fusion	**SJE**	**LatEm**	**SynC**	**SP**
w	51.7	47.4	48.8	**54.1**
w/o	29.9	34.9	35.2	**35.3**

Table 7: Comparison of SP method with SJE[24], LatEm[26] and SynC[7] in each semantic space, average per-class Top-1 accuracy(%) of unseen classes is reported based on the same data configurations, same images and semantic features in Dogs.

Semantic feature	**SJE**	**LatEm**	**SynC**	**SP**
att	N/A	N/A	N/A	N/A
w2v	19.6	22.6	27.6	**33.3**
glo	17.8	20.9	21.9	**33.4**
hie	24.3	25.2	31.1	**32.4**

4.8. SUN

SUN Attribute (SUN)[49] is the first large-scale scene attribute dataset. Because scene is greatly more complex than the specific object (e.g. bear, dog, or bird), so it is difficult to find the unsupervised source (e.g. wikipedia for w2v and glo) for precisely describing the scene semantics. Therefore, we only use att for implementing the experiment. Figure 7 indicates the experimental comparison of the different method in the attribute semantic space of SUN. Table 9 demonstrates the performance of the SP method is superior to that of other three methods. Specifically, the accuracy of SP can be enhanced by 11.5% than SJE, 10% than LatEm, and 4.8% than SynC.

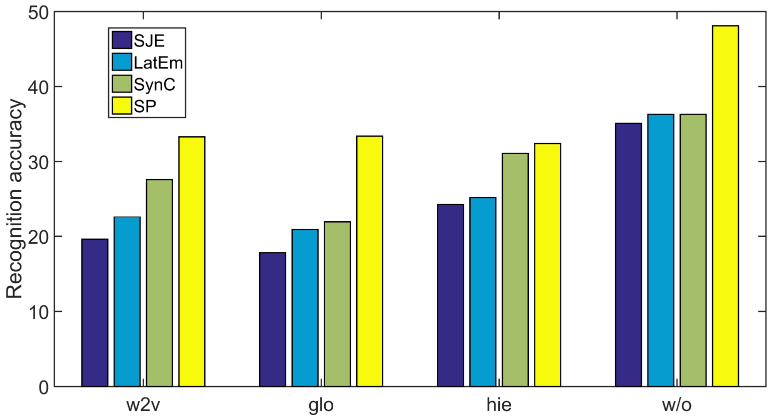

Figure 6: Comparison of SP method with SJE[24], LatEm[26] and SynC[7] in w2v, glo, hie and w/o (the fusion contains w2v, glo and hie), average per-class Top-1 accuracy of unseen classes is reported based on the same data configurations, same images and semantic features in Dogs.

4.9. Structure propagation with the iteration

The main idea of the proposed SP method shows three contents. In the first content, the manifold structure of images is considered for constructing the compatibility function between the class label and the visual feature. In the second content, the relationship between multi-manifold structures is found for booting the influence of the positive structure. In the last content, it is the most important to propagate the positive structure and fuse multi-manifold structures by the iteration computation. Therefore, we carry out the related experiment for evaluating the effect of the iteration on the structure evolution in AwA. The recognition accuracy can show the approximation degree of the class manifold structure. In other word, the better recognition accuracy is proportional to the more similar relationship between the reconstruction manifold structure and the intrinsic manifold structure of classes. Figure 8 demonstrates the recognition accuracy change with the iteration. In the beginning, the recognition accuracy rapidly

Table 8: Comparison of SP method with SJE[24], LatEm[26] and SynC[7] for multi-semantic fusion, average per-class Top-1 accuracy(%) of unseen classes is reported based on the same data configurations, same images and semantic features in Dogs. w: the fusion includes att, w2v, glo and hie, while w/o: the fusion contains w2v, glo and hie.

Fusion	**SJE**	**LatEm**	**SynC**	**SP**
w	N/A	N/A	N/A	N/A
w/o	35.1	36.3	36.3	**48.1**

Table 9: Comparison of SP method with SJE[24], LatEm[26] and SynC[7] in attribute semantic space, average per-class Top-1 accuracy(%) of unseen classes is reported based on the same data configurations, same images and semantic features in SUN.

Semantic feature	**SJE**	**LatEm**	**SynC**	**SP**
att	56.1	57.6	62.8	**67.6**

increases with the iteration, and then reaches a stable state. It means that structure propagation with the iteration can advance the recognition accuracy and finally obtain the best state.

4.10. Comparison with state-of-the-arts

In term of the image data utilization of unseen classes in testing, we can divide ZSL methods into two categories, which are inductive ZSL and transductive ZSL. Inductive ZSL methods can serially process unseen samples without the consideration of the underlying manifold structure in unseen samples[24] [26] [7] [23], while transductive ZSL can usually use the manifold structure of unseen samples to improve ZSL performance [21] [52] [45]. SP can find the structure of unseen classes in image feature space to enhance the transfer model between seen and unseen classes, so SP belongs to a transductive ZSL method. For a fair comparison, we use deep feature of images based on GoogleNet[50] in contrasting methods, which include our method, one transductive

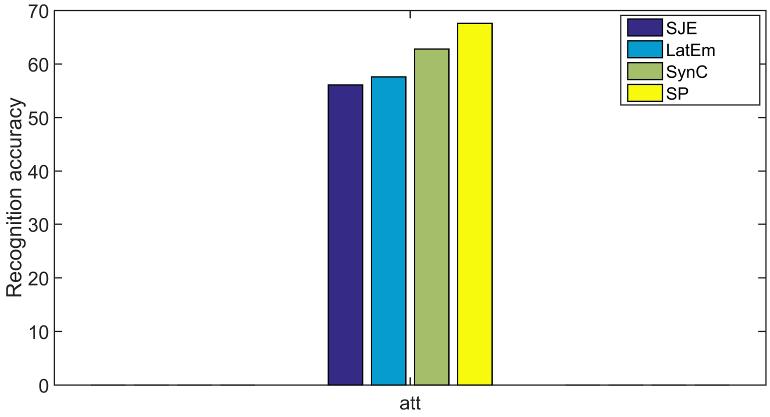

Figure 7: Comparison of SP method with SJE[24], LatEm[26] and SynC[7] in att, average per-class Top-1 accuracy of unseen classes is reported based on the same data configurations, same images and semantic features in SUN.

ZSL method (DMaP [45]), and three inductive ZSL methods(SJE[24], LatEm[26] and SynC[7]). To the best of our knowledge, these methods are state-of-the-art methods for ZSL. Table 10 shows their results for ZSL on three benchmark datasets. SP mostly outperforms the state-of-the-art methods except DMaP on CUB. DMaP focuses on the manifold structure consistency between the semantic representation and the image feature, and can better distinguish fine-grained classes. SP can complement the manifold structure between the semantic representation and the image feature, and better recognize coarse-grained classes. Therefore,integrating two ideas is expected to further improve the ZSL performance in future work.

4.11. Experimental result analysis

From the above mention, five methods for constructing the compatibility function have different consideration of the manifold structure. SJE can structure the output space by weighting the different output embedding, which can be associated with the

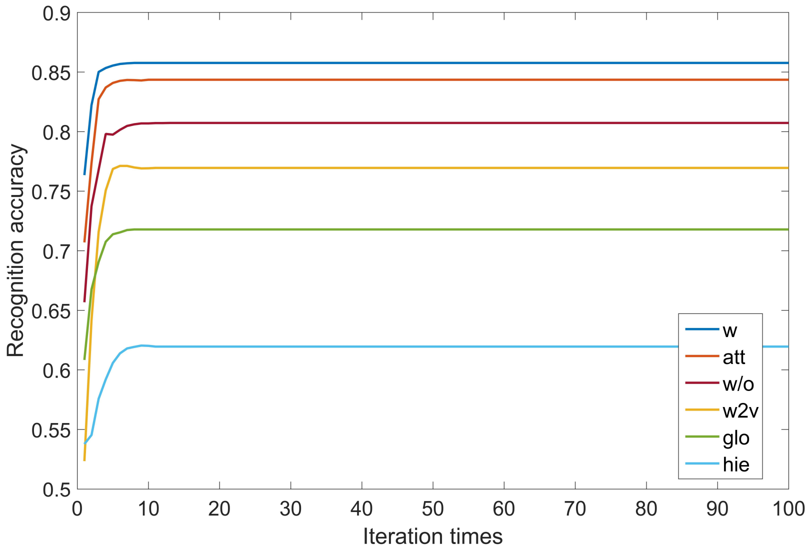

Figure 8: Average per-class Top-1 accuracy(%) of unseen classes is reported with structure propagation iteration times on AwA. w: the fusion includes att, w2v, glo and hie, while w/o: the fusion contains w2v, glo and hie

confidence contribution. LatEm try to find the structured model for making the overall piecewise linear function and can capture the flexible model of the latent space for fitting the unseen class. SynC can consider the manifold structure in semantic space for achieving optimal discriminative performance in the model space. DMaP can construct the manifold structure consistency between semantic representation and image feature. The proposed SP can take into consideration for optimizing the relationship of the manifold structure in semantic and image space, and enhance the positive structure propagation by iteration computation for ZSL. From the above experiments, we can attain the following observations.

- The semantic description have the different contribution for classifying unseen classes. The supervised attribute tend to obtain the better recognition performance than the unsupervised semantic representation (w2v, glo and hie) in AwA and CUB. In the unsupervised semantic representation, the recognition accuracy

Table 10: Comparison of SP method with state-of-the-art methods for ZSL, average per-class Top-1 accuracy (%) of unseen classes is reported based on the same data configurations. '+' indicates fusion operation.

Method	**Semantic feature**	**T/I**	**AwA**	**CUB**	**Dogs**
SJE	att	I	66.7	50.1	N/A
	w2v	I	51.2	28.4	19.6
LatEm	att	I	71.9	45.5	N/A
	w2v	I	61.1	31.8	22.6
SynC	att	I	69.3	47.5	N/A
	w2v	I	52.9	32.3	27.6
DMaP	att	T	74.9	**61.8**	N/A
	w2v	T	67.9	31.6	38.9
	att+w2v	T	78.6	59.6	N/A
SP	att	T	84.3	51.8	N/A
	w2v	T	77.4	32.5	33.3
	att+w2v	T	84.7	52.5	N/A
	att+w2v+glo+hie	T	**85.4**	54.1	N/A
	w2v+glo+hie	T	81.4	35.3	**48.1**

of w2v or glo is better than that of hie in in AwA and CUB, but the performance of hie is superior to that of w2v or glo in Dogs. This is mainly due to that the flexibility and uncertainty of the semantic representation in the unsupervised way.

- The performance of SP is better than that of other three methods, which are SJE, LatEm,and SynC. However, the performance improvement is different in the various datasets. The obvious improvement can be found in AwA, Dogs and

SUN, while the slight improvement can be shown in CUB. The main reason of this situation is related to whether or not effectively to propagate the positive structure in the optimization computation in term of data differences.

- SP emphasizes on the different manifold structure complement, while DMaP focuses on the various manifold structure consistency. Therefore, the performance of SP is superior to that of DMaP because the structure complementarity plays the important role for learning transfer model in AwA and Dogs, and the performance of DMaP is better than that of SP because the structure consistency is a key point for classifying unseen classes in CUB.
- SP performs better with the positive structure propagation. SP has demonstrated great promise in above experiments due to multi-manifold structure consideration and alternated optimization between the weight computation and the manifold structure estimation for ZSL.
- The proposed fusion method can attain the better performance than the non-fusion method because of appropriate complementing each other. w or w/o always performs better on AwA, CUB and Dogs.
- The most computational load involved in SP is for solving quadratic programming problem. Specifically,the complexity is $O(PU^3 + Pk^3)$ for P iteration times.

5. Conclusion

We have proposed a new ZSL method, which called Structure Propagation (SP). This method can not only directly model the relevance among the manifold structures in semantic and image space, but also dynamically propagate the positive structure by the crossing iteration. Specifically, the proposed SP method mainly includes four parts. First, nonlinear model constructs the mapping relationship between the class label and the visual image representation. Second, graph describes the relevance between seen classes and unseen classes in semantic or image space. Three, loss function indicates the constrains relationship of multi-manifold structure to balance the structure dependance. Last, structure propagation is implemented by the crossing iteration computation between phantom classes and weights solving. For evaluating the proposed SP, we carry out the experiment on AwA, CUB, Dogs and SUN. Experimental results show that SP can obtain the promising results for ZSL.

6. Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments that help improve the quality of this paper. Especially, The authors thank to Dr. Yongqin Xian from MPI for Informatics, who provided the data source to us. This work was supported by NSFC (Program No.61771386), Natural Science Basic Research Plan in Shaanxi Province of China (Program No.2016JM6045,No.2017JZ020).

References

References

[1] Y. Zhang, E. Zhang, W. Chen, Deep neural network for halftone image classification based on sparse auto-encoder, Engineering Applications of Artificial Intelligence 50 (C) (2016) 245–255.

[2] H. Larochelle, D. Erhan, Y. Bengio, Zero-data learning of new tasks, in: Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence(AAAI), 2008, pp. 646–651.

[3] X. Yu, Y. Aloimonos, Attribute-based transfer learning for object categorization with zero/one training example, in: European Conference on Computer Vision(ECCV), 2010, pp. 127–140.

[4] M. Rohrbach, M. Stark, B. Schiele, Evaluating knowledge transfer and zero-shot learning in a large-scale setting, in: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2011, pp. 1641–1648.

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

[6] S. Huang, M. Elhoseiny, A. Elgammal, D. Yang, Learning hypergraph-regularized attribute predictors, in: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2015, pp. 409–417.

[7] S. Changpinyo, W. L. Chao, B. Gong, F. Sha, Synthesized classifiers for zero-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2016, pp. 5327–5336.

[8] Y. Xian, B. Schiele, Z. Akata, Zero-shot learning-the good, the bad and the ugly, arXiv preprint arXiv:1703.04394.

[9] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 1778–1785.

[10] C. H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 951–958.

[11] D. Parikh, K. Grauman, Relative attributes, in: IEEE International Conference on Computer Vision(ICCV), 2011, pp. 503–510.

[12] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, T. Mikolov, Devise: a deep visual-semantic embedding model, in: Advances in Neural Information Processing Systems(NIPS), 2013, pp. 2121–2129.

[13] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.

[14] R. Socher, M. Ganjoo, H. Sridhar, O. Bastani, C. D. Manning, A. Y. Ng, Zero-shot learning through cross-modal transfer, in: Advances in Neural Information Processing Systems(NIPS), 2013, pp. 935–943.

[15] D. Jayaraman, K. Grauman, Zero-shot recognition with unreliable attributes, in: Advances in Neural Information Processing Systems(NIPS), 2014, pp. 3464–3472.

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

[17] X. Li, Y. Guo, D. Schuurmans, Semi-supervised zero-shot classification with label representation learning, in: IEEE International Conference on Computer Vision(ICCV), 2015, pp. 4211–4219.

[18] Z. Li, E. Gavves, T. Mensink, C. G. Snoek, Attributes make sense on segmented objects, in: European Conference on Computer Vision(ECCV), 2014, pp. 350–365.

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

[20] B. Romera-Paredes, P. H. Torr, An embarrassingly simple approach to zero-shot learning., in: International Conference on Machine Learning(ICML), 2015, pp. 2152–2161.

[21] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, Transductive multi-view zero-shot learning, IEEE transactions on pattern analysis and machine intelligence 37 (11) (2015) 2332–2345.

[22] Y. Yu, Z. Ji, X. Li, J. Guo, Z. Zhang, H. Ling, F. Wu, Transductive zero-shot learning with a self-training dictionary approach, arXiv preprint arXiv:1703.08893.

[23] Z. Zhang, V. Saligrama, Zero-shot learning via semantic similarity embedding, in: IEEE International Conference on Computer Vision(ICCV), 2015, pp. 4166–4174.

[24] Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele, Evaluation of output embeddings for fine-grained image classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2927–2936.

[25] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Label-embedding for image classification, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (7) (2016) 1425–1438.

[26] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, B. Schiele, Latent embeddings for zero-shot classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 69–77.

[27] G. Lin, K. Liao, B. Sun, Y. Chen, F. Zhao, Dynamic graph fusion label propagation for semi-supervised multi-modality classification, Pattern Recognition 68 (2017) 14–23.

[28] G. Lin, C. Fan, H. Zhu, Y. Miu, X. Kang, Visual feature coding based on heterogeneous structure fusion for image classification, Information Fusion 36 (2017) 275 – 283.

[29] G. Lin, G. Fan, X. Kang, E. Zhang, L. Yu, Heterogeneous feature structure fusion for classification, Pattern Recognition 53 (2016) 1 – 11.

[30] G. Lin, H. Zhu, X. Kang, C. Fan, E. Zhang, Feature structure fusion and its application, Information Fusion 20 (2014) 146 – 154.

[31] G. Lin, H. Zhu, X. Kang, Y. Miu, E. Zhang, Feature structure fusion modelling for classification, IET Image Processing 9 (10) (2015) 883–888.

[32] G. Lin, G. Fan, L. Yu, X. Kang, E. Zhang, Heterogeneous structure fusion for target recognition in infrared imagery, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2015, pp. 118–125.

[33] G. Lin, H. Zhu, X. Kang, C. Fan, E. Zhang, Multi-feature structure fusion of contours for unsupervised shape classification, Pattern Recognition Letters 34 (11) (2013) 1286 – 1290.

[34] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems(NIPS), 2013, pp. 3111–3119.

[35] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Conference on Empirical Methods in Natural Language Processing(EMNLP), 2014, pp. 1532–1543.

[36] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Label-embedding for attribute-based classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 819–826.

[37] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, S. F. Chang, Designing category-level attributes for discriminative visual recognition, in: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2013, pp. 771–778.

[38] M. Elhoseiny, B. Saleh, A. Elgammal, Write a classifier: Zero-shot learning using purely textual descriptions, in: IEEE International Conference on Computer Vision(ICCV), 2013, pp. 2584–2591.

[39] X. Li, Y. Guo, D. Schuurmans, Semi-supervised zero-shot classification with label representation learning, in: IEEE International Conference on Computer Vision(ICCV), 2016, pp. 4211–4219.

[40] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Label-embedding for image classification, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (7) (2016) 1425–1438.

[41] G. J. Qi, W. Liu, C. Aggarwal, T. S. Huang, Joint intermodal and intramodal label transfers for extremely rare or unseen classes, IEEE Transactions on Pattern Analysis and Machine Intelligence PP (99) (2016) 1–1. `doi:10.1109/TPAMI.2016.2587643`.

[42] Z. Zhang, V. Saligrama, Zero-shot learning via joint latent similarity embedding, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 6034–6042.

[43] Z. Fu, T. A. Xiang, E. Kodirov, S. Gong, Zero-shot object recognition by semantic manifold distance, in: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2015, pp. 2635–2644.

[44] T. Mensink, E. Gavves, C. G. M. Snoek, Costa: Co-occurrence statistics for zero-shot classification, in: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2014, pp. 2441–2448.

[45] Y. Li, D. Wang, H. Hu, Y. Lin, Y. Zhuang, Zero-shot recognition using dual visual-semantic mapping paths, arXiv preprint arXiv:1703.05002.

[46] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[47] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds200-2011 dataset, California Institute of Technology.

[48] J. Deng, J. Krause, L. Fei-Fei, Fine-grained crowdsourcing for fine-grained recognition, in: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2013, pp. 580–587.

[49] G. Patterson, C. Xu, H. Su, J. Hays, The sun attribute database: Beyond categories for deeper scene understanding, International Journal of Computer Vision 108 (1) (2014) 59–81.

[50] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[51] G. A. Miller, Wordnet: A lexical database for the english language, Contemporary Review 241 (1) (2002) 206–208.

[52] E. Kodirov, T. Xiang, Z. Fu, S. Gong, Unsupervised domain adaptation for zero-shot learning, in: IEEE International Conference on Computer Vision(ICCV), 2015, pp. 2452–2460.