Structure Fusion and Propagation for Zero-Shot Learning

Guangfeng Lin(✉), Yajun Chen, and Fan Zhao

Xi'an University of Technology, Xi'an 710048, Shaanxi Province,
People's Republic of China
{lgf78103,chenyajun,vcu}@xaut.edu.cn

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 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 fusion and propagation (SFP) for improving the performance of ZSL for classification. SFP 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 SFP can describe the constraint condition between the compatibility function and these manifold structures for balancing the influence of the structure fusion and propagation iteration. The SFP solution provides not only unseen class labels but also the relationship of two manifold structures that encodes the positive transfer in structure fusion and propagation. Experiments demonstrate that SFP can attain the promising results on the AwA, CUB, Dogs and SUN datasets.

Keywords: Structure fusion and propagation · Manifold structure
Zero-shot learning · Transfer learning

1 Introduction

Although deep learning [32] depending on large-scale labeled data training has been generally used for visual recognition [31], a daunting challenge still exists to recognize visual object "in the wild". In fact, in specific applications it is

Supported by NSFC (Program No. 61771386, Program No. 61671376 and Program No. 61671374), Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2016JM6045, Program No. 2017JZ020).

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11258, pp. 465–477, 2018.
https://doi.org/10.1007/978-3-030-03338-5_39

impossible to collect all class data for training deep model, so training (seen classes) and testing classes(unseen classes) are often disjoint. The main idea of ZSL is to handle this problem by exploiting the transfer model from the redundant relevance of the semantic description. To recognize unseen classes from seen classes, ZSL needs face to two challenges [3]. 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.

ZSL can bridge the gap among the different domains to recognize unseen class objects by semantic embedding of class labels. These semantic embeddings can come from vision (attributes [11]) and language information (text [25]) by the manual annotation, machine learning [29]or data mining [5]. In term of the transformation relationship of different embedding, recent ZSL methods mainly fall into linear embedding, nonlinear embedding and similarity embedding. Linear embedding [1,2,7,13,24] implements the linear transformation method among different embedding spaces for learning the relevance between unseen class objects and class labels. Nonlinear embedding [23,25,28] can realize the nonlinear mapping of the embedding space for building the compatibility function or classifier, which can be learned by deep networks [14,30]. Similarity embedding [3,9,15,19,33] builds the classifier by the similarity metrics, which mostly include structure learning or class-wise similarities. 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.

In this paper, our motivation is inspired by structure fusion [16–18] 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. 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. Model space corresponding to visual appearances is the jointed projection space of semantic space and image space, and can preserve the respective manifold structure. Figure 1 illustrates the idea of the proposed method conceptually. SFP considers not only semantic and image structures but also the positive structure propagation for ameliorating unseen objects classification, while SynC [3] only focus on manifold structure in semantic space for combining the base classifier in ZSL.

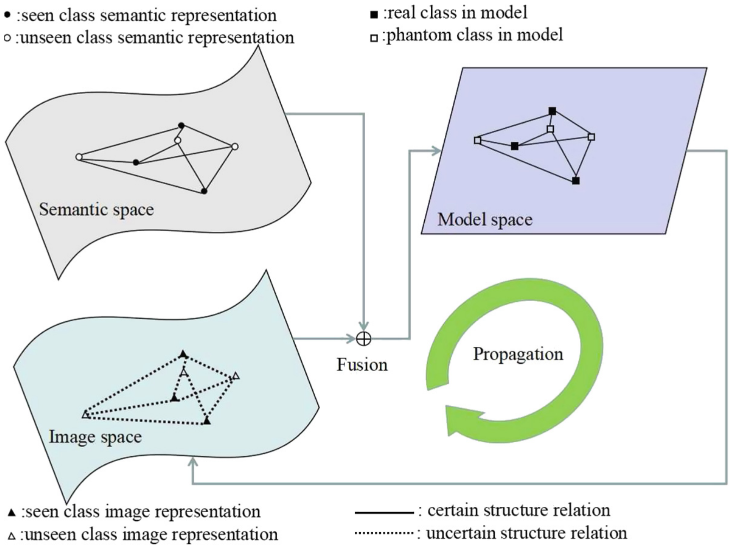
● :seen class semantic representation
○ :unseen class semantic representation
■ :real class in model
□ :phantom class in model
Semantic space
Model space
Image space
Fusion
Propagation
▲ :seen class image representation
△ :unseen class image representation
———— : certain structure relation
··············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

Fig. 1. The illustration of structure fusion and propagation for zero-shot learning. 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 [3].) and real object classes corresponding to all classes in model space.

In our main contribution, a novel idea have tow aspects 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, one constructs the projected manifold structure for real and phantom class in model space, another constrains the compatibility function and the relationship of the manifold structure for the positive structure propagation.

2 Structure Fusion and 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 Table 1) 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.

2.1 Classification Model and Manifold Structure

We construct a pair-wise linear classifier [3] 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 graphs, 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); $G_x{<}V_x, E_x{>}$ 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). Therefore, the similarity of semantic and image space is respectively regarded as the weight between nodes, which can be defined as following.

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

here, b_s and x_s are respectively the semantic and image representation (the detail is described in Table 1) of the seen class s, while b_u and x_u are respectively the semantic and image representation of the unseen class u. $w_{su}^{(b)}$ and $w_{su}^{(x)}$ are respectively the weight (the similarity) between the seen class s and the unseen class u in semantic and image representation space. $d(b_s, b_u)$ and $d(x_s, x_u)$ are respectively the distance metric [3] of each space, and can be defined as following.

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

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 another 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.) $\Sigma_x = \sigma_x I$ 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, while 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 graphs 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.

$$(a_c, v_u, \boldsymbol{\beta}) = \arg \min_{a_c, v_u, \boldsymbol{\beta}} \|a_c - \sum_{u=1}^{U} \boldsymbol{\beta}^T \left[w_{su}^{(x)} \ w_{su}^{(b)} \right]^T v_u - \sum_{s=1}^{S} \boldsymbol{\gamma}^T \left[w_{ss}^{(x)} \ w_{ss}^{(b)} \right]^T v_s \|_2^2,$$
$$s.t. \quad \boldsymbol{\beta}^T \mathbf{1} = 1, \boldsymbol{\gamma}^T \mathbf{1} = 1, 0 \leq \beta_i \leq 1, 0 \leq \gamma_i \leq 1 \quad (i = 1, 2) \tag{4}$$

here, $\boldsymbol{\beta} = \left[\beta_1 \ \beta_2\right]^T$, $\boldsymbol{\gamma} = \left[\gamma_1 \ \gamma_2\right]^T$, and $\mathbf{1} = \left[1 \ 1\right]^T$. Because no connection exists between unseen classes or seen classes in tow bipartite graphs, $w_{ss}^{(b)} = 0$ and $w_{ss}^{(x)} = 0$. The analytical solution of (4) can find the relation between a_c and v_u.

$$a_c = \sum_{u=1}^{U} \boldsymbol{\beta}^T \left[w_{su}^{(x)} \ w_{su}^{(b)} \right]^T v_u,$$
$$s.t. \quad \boldsymbol{\beta}^T \mathbf{1} = 1, 0 \leq \beta_i \leq 1 \quad (i = 1, 2) \tag{5}$$

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

2.2 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 [3].

$$\begin{aligned}(v_1,...,v_U,\boldsymbol{\beta}) = &\arg\min_{v_1,...,v_U,\boldsymbol{\beta}} \sum_{c=1}^{S}\sum_{n=1}^{N} \ell(x_n, \mathbb{I}_{y_n,c}, a_c) \\ &+ \frac{\lambda}{2}\sum_{c=1}^{S}\|a_c\|_2^2 + \frac{\gamma}{2}\|\beta_1 W^x - \beta_2 W^b\|_2^2, \\ &s.t. \quad a_c = \sum_{u=1}^{U} \boldsymbol{\beta}^T \left[w_{su}^{(x)}\ w_{su}^{(b)}\right]^T v_u, \\ &\boldsymbol{\beta}^T\mathbf{1} = 1, 0 \le \beta_i \le 1 \quad (i = 1, 2)\end{aligned} \tag{6}$$

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 first term of formula (6) is the squared hinge loss, which can be defined as $\ell(x_n, \mathbb{I}_{y_n,c}, a_c) = \max(0, 1 - \mathbb{I}_{y_n,c} a_c x_n)$. $\mathbb{I}_{y_n,c} \in \{-1, 1\}$ determines whether or not $y_n = c$. The second term of formula (6) 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. The third term of formula (6) 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 (6) with respect to $\{v_u\}_{u=1}^{U}$ and $\boldsymbol{\beta}$ by solving the quadratic programming problem.

To depict the whole process of the structure fusion and propagation mechanism, we show the pseudo code of the proposed SFP algorithm in Algorithm 1.

Algorithm 1. The pseudo code of the SFP 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)
Computes the similarity matrix $W_{(b)}$ on the semantic representation by (2)
Setting the similarity matrix $W_{(x)}$ to zero matrix on the image representation
for $1 < t < P$ do
    Solving $\{v_u\}_{u=1}^{U}$ and $\boldsymbol{\beta}$ by alternately optimizing (6)
    Computing $a_c$ according to (5)
    Computing $\hat{y}$ by (1) and obtaining the class label $y_t^*$ of the unseen class corresponding to the semantic class
    Computing the mean value of each image class as the image class representation $x_s$ and $x_u$
    Computing and updating the similarity matrix $W_{(x)}$ on the image representation by (2)
end for
```

2.3 Complexity Analysis

Formula (6) can be solved by alternately quadratic programming, which of the complexity includes two parts. In the first part, when $\boldsymbol{\beta}$ is fixed, formula (6) is related to $\{v_u\}_{u=1}^{U}$ of a quadratic programming problem, which of the complexity is $O(U^3)$ for the worst. In the second part, while $\{v_u\}_{u=1}^{U}$ is fixed, formula (6) is corresponding to $\boldsymbol{\beta}$ of a quadratic programming problem, which of the complexity is $O(k^3)$ (k is the dimension of $\boldsymbol{\beta}$) for the worst. Given the proposed algorithm SFP needs P iterations, it's complexity is $O(PU^3 + Pk^3)$.

3 Experiment

3.1 Datasets

For evaluating the proposed algorithm SFP[1], we carry out the experiment in four challenging datasets, which are Animals with Attributes (AwA) [12], CUB-200-2011 Birds (CUB) [27], Stanford Dogs (Dogs) [4], and SUN Attribute (SUN) [21]. 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 [6], w2v [20], glo [22] and hie [1], while Dogs is represented by w2v [20], glo [22] and hie [1]. SUN is only depicted by att [6]. Table 1 provides the statistics and the extracted features for these datasets. In addition, for conveniently comparing with the state-of-art methods, we adopt image feature provided by [1].

Table 1. 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 [26]/1024
CUB	150	50	11786	att/312, w2v/400, glo/400, hie/about 200	Deep feature based on GoogleNet [26]/1024
Dogs	85	28	19499	N/A, w2v/400, glo/400, hie/about 200	Deep feature based on GoogleNet [26]/1024
SUN	645	72	14340	att/102, N/A, N/A, N/A	Deep feature based on GoogleNet [26]/1024

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

3.2 Comparison with the Baseline Methods

In this paper, there are three methods as the baseline for comparing with the proposed SFP method because of the semantic structure mining. The first method is structured joint embedding (SJE) [1], which can build the bilinear compatibility function with consideration of the structured output space for predicting the label of the unseen class. The second method is latent embedding model (LatEm) [28],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) [3], which can make nonlinear compatibility function with manifold structure in semantic space for combining the base classifier in ZSL. Table 2 shows the performance of the structure fusion and propagation (the proposed SFP method) greatly outperforms that of other three methods.

3.3 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 $\lambda, \gamma, \sigma_b$ and σ_x (respectively are in formula (3) in formula (6)). 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 another 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.

3.4 Structure Fusion and Propagation with the Iteration

The main idea of the proposed SFP 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 2 demonstrates the recognition accuracy change with the iteration. In the beginning, the recognition accuracy rapidly increases with the iteration, and then reaches a stable state. It means that structure fusion and propagation with the iteration can advance the recognition accuracy and finally obtain the best state.

Table 2. Comparison of SFP method with SJE [1], LatEm [28] and SynC [3] 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. w: the fusion includes att, w2v, glo and hie, while w/o: the fusion contains w2v, glo and hie.

Datasets	Semantic feature	SJE	LatEm	SynC	SFP
AwA	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**
	w	73.9	76.1	78.0	**85.4**
	w/o	60.1	66.2	69.1	**81.4**
CUB	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**
	w	51.7	47.4	48.8	**54.1**
	w/o	29.9	34.9	35.2	**35.3**
Dogs	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**
	w	*N/A*	*N/A*	*N/A*	*N/A*
	w/o	35.1	36.3	36.3	**48.1**
SUN	att	56.1	57.6	62.8	**67.6**

3.5 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 [1,3,28,33], while transductive ZSL can usually use the manifold structure of unseen samples to improve ZSL performance [8,10,15]. SFP can find the structure of unseen classes in image feature space to enhance the transfer model between seen and unseen classes, so SFP belongs to a transductive ZSL method. For a fair comparison, we use deep feature of images based on GoogleNet [26] in contrasting methods, which include our method, one transductive ZSL method (DMaP [15]), and three inductive ZSL methods (SJE [1], LatEm [28] and SynC [3]). To the best of our knowledge, these methods are state-of-the-art methods for ZSL. Table 3 shows their results for ZSL on three benchmark datasets. SFP mostly outperforms the state-of-the-art methods except DMaP on CUB. DMaP focuses on the manifold

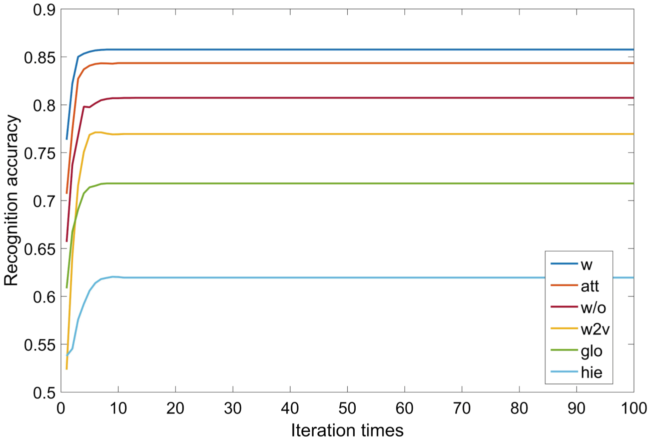

Fig. 2. Average per-class Top-1 accuracy (%) of unseen classes is reported with structure fusion and propagation iteration times on AwA. w: the fusion includes att, w2v, glo and hie, while w/o: the fusion contains w2v, glo and hie

structure consistency between the semantic representation and the image feature, and can better distinguish fine-grained classes. SFP 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.

3.6 Experimental Result Analysis

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 of w2v or glo is better than that of hie in AwA and CUB, but the performance of hie is superior to that of w2v or glo in Dogs. This is mainly due to the flexibility and uncertainty of the semantic representation in the unsupervised way.
- The performance of SFP 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.
- SFP emphasizes on the different manifold structure complement, while DMaP focuses on the various manifold structure consistency. Therefore, the performance of SFP is superior to that of DMaP because the structure complementarity plays the important role for learning transfer model in AwA and

Table 3. Comparison of SFP 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
SFP	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**

Dogs, and the performance of DMaP is better than that of SFP because the structure consistency is a key point for classifying unseen classes in CUB.
- SFP performs better with the positive structure fusion and propagation. SFP 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.

4 Conclusion

We have proposed a new ZSL method, which called structure fusion and propagation (SFP). 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 SFP 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 fusion and propagation is implemented by the crossing iteration computation between phantom classes and weights solving. For evaluating the proposed SFP, we carry out the experiment on AwA, CUB, Dogs and SUN. Experimental results show that SFP can obtain the promising results for ZSL.

References

1. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2927–2936 (2015)
2. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for image classification. IEEE Trans. Pattern Anal. Mach. Intell. **38**(7), 1425–1438 (2016)
3. Changpinyo, S., Chao, W.L., Gong, B., Sha, F.: Synthesized classifiers for zero-shot learning. In: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 5327–5336 (2016)
4. Deng, J., Krause, J., Fei-Fei, L.: Fine-grained crowdsourcing for fine-grained recognition. In: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 580–587 (2013)
5. Elhoseiny, M., Saleh, B., Elgammal, A.: Write a classifier: zero-shot learning using purely textual descriptions. In: IEEE International Conference on Computer Vision(ICCV), pp. 2584–2591 (2013)
6. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1778–1785 (2009)
7. Frome, A., et al.: DeViSE: a deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems (NIPS), pp. 2121–2129 (2013)
8. Fu, Y., Hospedales, T.M., Xiang, T., Gong, S.: Transductive multi-view zero-shot learning. IEEE Trans. Pattern Anal. Mach. Intell. **37**(11), 2332–2345 (2015)
9. Fu, Z., Xiang, T.A., Kodirov, E., Gong, S.: Zero-shot object recognition by semantic manifold distance. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2635–2644 (2015)
10. Kodirov, E., Xiang, T., Fu, Z., Gong, S.: Unsupervised domain adaptation for zero-shot learning. In: IEEE International Conference on Computer Vision (ICCV), pp. 2452–2460 (2015)
11. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 951–958 (2009)
12. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE Trans. Pattern Anal. Mach. Intell. **36**(3), 453–465 (2014)
13. Li, X., Guo, Y., Schuurmans, D.: Semi-supervised zero-shot classification with label representation learning. In: IEEE International Conference on Computer Vision (ICCV), pp. 4211–4219 (2016)
14. Li, Y., Zhang, J., Zhang, J., Huang, K.: Discriminative learning of latent features for zero-shot recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7463–7471 (2018)
15. Li, Y., Wang, D., Hu, H., Lin, Y., Zhuang, Y.: Zero-shot recognition using dual visual-semantic mapping paths. arXiv preprint arXiv:1703.05002 (2017)
16. Lin, G., Fan, C., Zhu, H., Miu, Y., Kang, X.: Visual feature coding based on heterogeneous structure fusion for image classification. Inf. Fusion **36**, 275–283 (2017)

17. Lin, G., Fan, G., Kang, X., Zhang, E., Yu, L.: Heterogeneous feature structure fusion for classification. Pattern Recognit. **53**, 1–11 (2016)
18. Lin, G., Liao, K., Sun, B., Chen, Y., Zhao, F.: Dynamic graph fusion label propagation for semi-supervised multi-modality classification. Pattern Recognit. **68**, 14–23 (2017)
19. Mensink, T., Gavves, E., Snoek, C.G.M.: Costa: co-occurrence statistics for zero-shot classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2441–2448 (2014)
20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (NIPS), pp. 3111–3119 (2013)
21. Patterson, G., Xu, C., Su, H., Hays, J.: The sun attribute database: beyond categories for deeper scene understanding. Int. J. Comput. Vis. **108**(1), 59–81 (2014)
22. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
23. Qi, G.J., Liu, W., Aggarwal, C., Huang, T.S.: Joint intermodal and intramodal label transfers for extremely rare or unseen classes. IEEE Trans. Pattern Anal. Mach. Intell. **PP**(99), 1 (2016). https://doi.org/10.1109/TPAMI.2016.2587643
24. Romera-Paredes, B., Torr, P.H.: An embarrassingly simple approach to zero-shot learning. In: International Conference on Machine Learning (ICML), pp. 2152–2161 (2015)
25. Socher, R., Ganjoo, M., Sridhar, H., Bastani, O., Manning, C.D., Ng, A.Y.: Zero-shot learning through cross-modal transfer. In: Advances in Neural Information Processing Systems (NIPS), pp. 935–943 (2013)
26. Szegedy, C., et al.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9 (2015)
27. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds200-2011 dataset. California Institute of Technology (2011)
28. Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., Schiele, B.: Latent embeddings for zero-shot classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 69–77 (2016)
29. Yu, F.X., Cao, L., Feris, R.S., Smith, J.R., Chang, S.F.: Designing category-level attributes for discriminative visual recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 771–778 (2013)
30. Zhang, C., Peng, Y.: Visual data synthesis via GAN for zero-shot video classification. arXiv preprint arXiv:1804.10073 (2018)
31. Zhang, E., Chen, W., Zhang, Z., Zhang, Y.: Local surface geometric feature for 3D human action recognition. Neurocomputing **208**, 281–289 (2016)
32. Zhang, Y., Zhang, E., Chen, W.: Deep neural network for halftone image classification based on sparse auto-encoder. Eng. Appl. Artif. Intell. **50**, 245–255 (2016)
33. Zhang, Z., Saligrama, V.: Zero-shot learning via joint latent similarity embedding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6034–6042 (2016)