Structurally Constrained Correlation Transfer for Zero-shot Learning

Yu Chen, Yuehan Xiong, Xing Gao, Hongkai Xiong
Department of Electronic Engineering, *Shanghai Jiao Tong University*, *Shanghai, China*
{chen_yu_, xiongyuehan, william-g, xionghongkai}@sjtu.edu.cn

***Abstract*—Given a set of labeled data with semantic descriptions, zero-shot learning aims at recognizing objects from unseen classes, where no instances of the classes are used during training. Most existing methods solve this problem via embedding images and labels into an embedding space and computing similarity across different information sources. However, the similarity calculation could be unreliable when the training and testing data distributions are inconsistent. In this paper, we propose a novel zero-shot learning model that forms a neighborhood-preserving structure in the semantic embedding space and utilize it to predict classifiers for unseen classes. By constructing a locally connected graph for class embeddings, we exploit the structural constraint of embeddings of similar classes and retain the global structure in the semantic embedding space to obtain an effective representation of semantic information. Experiment results on three benchmark datasets demonstrate that the proposed method generates effective semantic representations and out-performs state-of-the-art methods.**

***Index Terms*—object recognition, zero-shot learning, semantic embeddings, manifold learning**

I. Introduction

Supervised learning methods have achieved significant progress in multiple fields, where performance are highly dependent on large-scale labeled data that are not always available. In contrast, a previously unseen class can be recognized when given a set of labeled data under the circumstances of zero-shot learning (ZSL). To achieve this goal, an effective description of unseen data imposes a significant impact on this task. A practical way to describe unseen data is semantic embedding, and a large number of semantic embedding methods have been explored, such as attributes, word vectors and attempts that integrate multiple types of embeddings.

The current ZSL methods generally fall into four categories [1]: learning attribute classifiers-based, learning linear compatibility-based, learning nonlinear compatibility-based and hybrid models-based. Attribute classifiers-based methods attempt to learn classifiers for each attribute [2], but deviation may be introduced in the process of establishing classifiers. Linear compatibility-based and nonlinear compatibility-based methods directly measure the correlation between images and semantic embeddings, but shift may occur with inconsistent distributions of the training data and testing data. Thereafter, hybrid models-based methods are proposed, which regard class embeddings as the combination of seen class proportions. In [3], semantic embeddings for unseen classes are synthesized via the combination of semantic embeddings of seen classes weighted by their corresponding probabilities from pre-trained classifiers. Nevertheless, training classifiers in advance are time-consuming and inefficient. Besides, a recent method [4] achieving the state-of-the-art performance in ZSL utilizes phantom classes to transfer knowledge between semantic embeddings and classifiers, and thereby new classifiers are synthesized via the convex combination of phantom classes given the semantic embeddings. However, it ignores the local structure among semantic embeddings, which might provide more information to synthesize classifiers precisely.

In this paper, we propose a ZSL framework that assumes and exploits more structural relations in the semantic embedding space. The main idea is to explore the intuition that semantic representations from similar classes will be projected into the neighbor locations in the embedding space, which would help to predict classifiers for unseen classes. Specially, we leverage structural relations by taking the connection of neighbor embeddings of similar classes into consideration and propose to predict new classifiers through constructing a locally connected graph for unseen classes. While exploring the local relationship and retaining the global structure, the proposed method strengthens the effect of neighbor embeddings and obtain a more effective representation of semantic information. Fig. 1 illustrates our approach conceptually and experiment results on three benchmark datasets demonstrate the effectiveness of our method.

II. Structurally Constrained Correlation Transfer

A. Problem Formulation

Suppose there are n_S image-label pairs $\{(x_i, y_i)\}_{i=1}^{n_S}$ from S seen classes, x_i is the ith image and y_i is its label. The set of images and the set of labels in the seen classes are denoted as X_S and Y_S, respectively. And there are n_U unlabeled images $\{(x_j)\}_{j=1}^{n_U}$ from U unseen classes. The set of images and the set of labels in the unseen classes are denoted as X_U and Y_U, respectively. Here, we employ the semantic embeddings which are associated with Y_U as the side information. So the ZSL task is to obtain the correspondence $X_U \rightarrow Y_U$, where Y_U is disjoint from Y_S.

We consider predicting classifiers to identify unseen classes based on the knowledge acquired from seen classes. To achieve this goal, a model space consisting of classifiers trained with labeled data is constructed. Furthermore, to better exploit the class information, a semantic space is also constructed, which consists of class embeddings such as attributes or word

978-1-5386-4458-4/18/$31.00 ©2018 IEEE

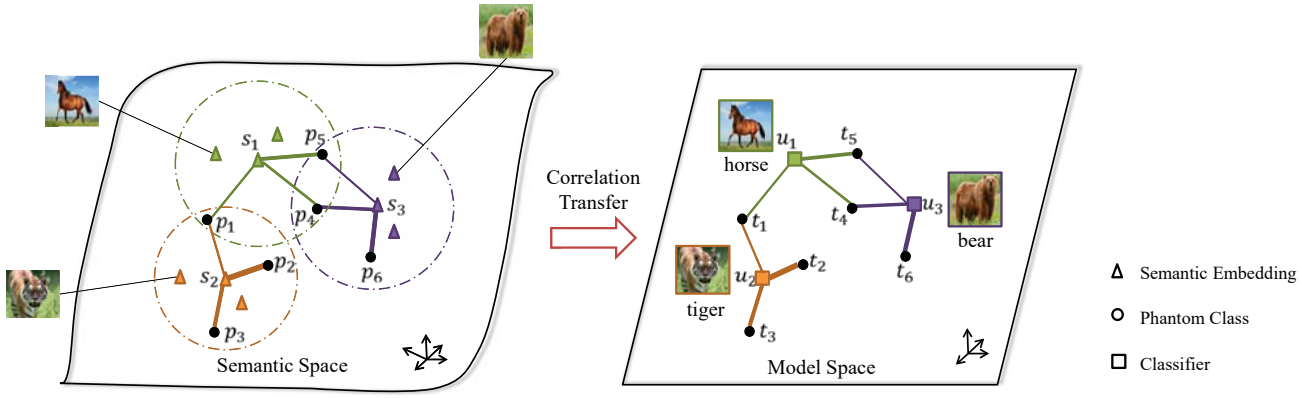


Fig. 1. The structurally constrained correlation transfer method for zero-shot learning. Suppose there exist a semantic space and a model space lying on the manifold, where the former consists of attributes and the latter consists of linear classifiers trained with labeled data. Each semantic embedding (**s**) is represented by its neighbor phantom classes (**p**). By forming and projecting the locally-connected class correlation from the semantic space to the model space via phantom classes (**p** and **t**), structural constraints in the semantic space can be transferred into the model space. Our method can predict classifiers (**u**) for unseen classes when given the semantic representations.

vectors. To project the correlation from the semantic space into the model space, each space is set on a manifold and phantom classes are introduced to transfer knowledge between two spaces. Considering the local structure in the semantic space, we propose to impose structural constraint on embeddings of similar classes, thus the semantic descriptions are constrained to predict classifiers for unseen data. Therefore, by viewing each semantic embedding and its neighbor phantom classes as vertices, a bipartite graph can be formed in the semantic space. The similarity between a semantic embedding and a phantom class can be defined as the weight of the corresponding graph edge. Considering the projection of the phantom classes from the semantic space to the model space as a nonlinear dimensionality reduction problem, the locally linear embedding [5] is modified to solve this problem.

III. Graph Construction in the Semantic Space

In this section, we propose to construct a graph for semantic embeddings of classes considering the embeddings of similar classes. Since attributes have shown to be superior to word vectors of class names [6], we consider using attributes as the semantic embeddings. Suppose each class i has a coordinate s_i and all classes live on a manifold in the semantic space, where $i \in \{1, 2, ..., S+U\}$. Besides, each classifier associated with class i has a coordinate u_i and all classifiers live on a manifold in the model space. Furthermore, R phantom classes are introduced to be viewed as the bridge between the semantic space and the model space. The coordinate of each phantom class is represented as p_m $(m = 1, ..., M)$ in the semantic space and its corresponding classifier is represented as t_m in the model space. We presume that each semantic embedding and its neighbor R_n phantom classes are lie on a locally linear patch of the manifold in the semantic space. This assumption allows us to reconstruct the nonlinear structure in the model space from the bipartite graph constructed in the semantic space. Inspired by [7], the weight between a real class s_i and a phantom class p_m in the semantic space is defined as:

$$w_{mi} = \frac{exp(-||p_m - s_i||^2)}{\sum_{k=1}^{R_n} exp(-||p_k - s_i||^2)} \tag{1}$$

where the scaled squared Euclidean distance are used as the similarity measurement. We also explore Manhattan distance and Chebyshev distance in the experiment, but neither of them can describe the distance on manifold appropriately. As shown in (1), the introduction of structurally constrained correlation enables us to project a more informative nonlinear structure from the semantic space to the model space. By considering embeddings of similar classes, the proposed method formulates locally connected graph and generates more discriminative classifiers with the constructed graph. In contrast, in previous work each embedded class is represented as the convex combination of phantom classes in the semantic space regardless of its neighboring embedding structure, which dampens the effect of neighboring embeddings.

A. Predict Classifiers in the Model Space

In the model space, each classifier u_i is represented by R_n neighbor phantom classes t_m via:

$$u_i = \sum_{m=1}^{R_n} w_{mi} t_m, \ \forall i \in \{1, 2, ..., S+U\} \tag{2}$$

where the weight w_{mi} is obtained from the constructed graph in the semantic space.

Substituting (1) into (2), we could obatin the classifier u_i of an unseen image, whose semantic embedding is s_i. An overview of the proposed method is depicted as in Fig. 1.

B. Parameter Learning

Since each phantom class p_m could be represented by its neighbor semantic embeddings as

$$p_m = \sum_{i=1}^{R_n} \mu_{mi} s_i, \ \forall i \in \{1, 2, ..., S+U\} \tag{3}$$

Hence we minimize the following objective function:

$$\min_{\{t_m\}_{m=1}^{R},\{\mu_{mi}\}_{m,i=1}^{R,S}} \sum_{i=1}^{S}\sum_{k=1}^{R_n} l(x_k, \zeta_{y_k,i}; u_i) + \lambda \sum_{i=1}^{S} \|u_i\|_2^2 + \gamma \sum_{m,i=1}^{R,S} |\mu_{mi}|,$$
$$s.t. \quad u_i = \sum_{m=1}^{R_n} w_{mi} t_m, \ \forall i \in \{1, ..., S\}$$

where one-vs-rest strategy is employed and squared hinge loss $l(x, y; u) = \max(0, 1 - yu^{\mathrm{T}}x)^2$ is used to ensure that classifiers focus on the overall classification error. The indicator $\zeta_{y_k,i} \in \{-1, 1\}$ represents whether $y_k = i$ or not.

Since the objective function is not convex for $\{t_m\}_{m=1}^{R}$ and $\{\mu_{mi}\}_{m,i=1}^{R,S}$, we deploy an alternating optimization to solve it. Furthermore, we test our method with the structured loss when considering the class relatedness based on the Crammer-Singer multi-class SVM loss [8], where

$$l_{struct} = \max(0, \max_{i \in S - \{y_k\}} ||s_i - s_{y_k}||_2^2 + {u_i}^{\mathrm{T}} x_k - u_{y_k}^{\mathrm{T}} x_k) \quad (4)$$

In the training phase, we first calculate the R_n nearest neighbors for each semantic embedding, then a local connection graph is formed. The phantom classes in the semantic space are projected into the model space. In the predicting phase, the semantic embeddings of each unseen class are represented by its R_n nearest phantom classes in the semantic space. When projecting the phantom classes and retaining the weighted graph from the the semantic space to the model space, new classifiers are synthesized based on (2) and identify unseen classes.

Since [5] has demonstrated that overlapping local neighborhoods can provide information about global geometry when they are collectively analyzed, our method exploits the local distributions of the semantic embeddings while retaining the global structure in the semantic space. When the local semantic representations are projected into the model space, the global nonlinear structure is recovered from locally linear connections.

IV. Experiments

This section presents an assessment to verify the effectiveness of our model. We compare our method with the state-of-the-art model [4], [9] and several representative methods [2], [3], [6], [10]. The assessment is conducted on the public data for consistency with prior methods.

A. Datasets and Settings

1) Training and Testing Set: We test our approach on three benchmark datasets. The first dataset is Animal with Attributes (AwA) [2], which contains 85 binary attributes and 30,475 images from 50 classes. The second dataset is CUB-200-2011 Birds (CUB) [11], which contains 312 attributes and 11,788 images from 200 bird classes. The third dataset is SUN Attribute (SUN) [12], which contains 102 attributes and 14,340 images from 717 scene categories. Table I shows the statistics and split methods of these datasets.

TABLE I
STATISTICS OF THE THREE DATASETS USED IN THE EXPERIMENTS. WE FOLLOW THE PRESCRIBED SPLIT IN [2] FOR AWA, THE PRESCRIBED SPLIT (4 SPLITS FOR CUB AND 10 SPLITS FOR SUN) IN [4] AND REPORT THE AVERAGE RESULTS.

Dataset	# seen classes	# unseen classes	# total images	# attributes
AwA	40	10	30,475	85
CUB	150	50	11,788	312
SUN	645/646	72/71	14,340	102

TABLE II
CLASSIFICATION ACCURACIES OF THE THREE DATASETS USED IN THE EXPERIMENTS IN (%). FOR SYNC [4], WE CITE RESULTS FROM OUR IMPLEMETATION. RESULTS WITH $'.^{*}'$ ARE GENERATED WITH PARTICULAR SPLITTING PROCEDURES.

Methods	AwA	CUB	SUN
DAP [2]	41.4	-	22.2
IAP [2]	42.2	-	18.0
ConSE [3]	63.3	36.2	51.9
COSTA [10]	61.8	40.8	47.9
SJE [6]	66.7	50.1^{*}	-
AHLE [9]	49.4	27.3^{*}	-
$\mathrm{SynC}^{\mathrm{ovo}}$ [4]	69.7	49.8	62.1
$\mathrm{SynC}^{\mathrm{struct}}$ [4]	72.9	**53.0**	62.4
$Ours^{\mathrm{ovo}}$	72.3	47.1	62.4
$Ours^{\mathrm{struct}}$	**74.6**	48.5	**62.7**

TABLE III
THE VALUES OF NEIGHBOR PHANTOM CLASSES R_n FOR DIFFERENT DATASETS. R IS THE NUMBER OF PHANTOM CLASSES.

Datasets	AwA	CUB	SUN
R_n (ovo)	0.4R	0.46R	0.713R
R_n (struct)	0.4R	0.45R	0.748R

2) Semantic space: We experiment with the features extracted with AlexNet [13] for images from AwA and CUB, and features extracted with GoogLeNet [14] by Caffe package [15] for images from SUN.

3) Performance Metric: The normalized multi-class classification accuracy is used as the evaluation protocol as in existing works.

B. Implementation Details

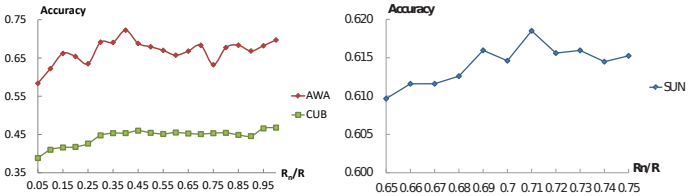
We obtain the class-attribute representation by averaging the attribute representations of the images in the same class. All variables are initialized randomly and hyper-parameters are selected based on the training data for each dataset via cross-validation strategy. Note that the cross-validation strategy splits the classes instead of splitting training data according to [4]. Besides, the distance between the unconnected vertices is set as ∞. Furthermore, the numbers of neighbor phantom classes R_n are set as the fractional multiples of the number of phantom classes R at the beginning, then we gradually narrow down the range based on the accuracy change to obtain more accurate results.

C. Results

In order to validate the effectiveness of the proposed method and compare it with representative methods, we conduct the following three experiments.

1) Comparision with Representative Models: In order to study the performance of the proposed model, we compare the ZSL classification accuracy of our method with seven representive approaches, as shown in Table II. We denote our method as $Ours^{\text{ovo}}$ (one-vs-others) and $Ours^{\text{struct}}$ (structured loss) when considering two different strategies. We observe that our method achieves better performance compared with the published methods in most scenarios due to the consideration of structural constraint of semantic embeddings. The best classification accuracies of our method on three datasets are 74.6%, 48.5% and 62.7%, which outperforms all models on AwA and SUN under two different loss constraints. We also notice that the performance of the proposed method is less effective on CUB, which has more attributes compared with the other two datasets as listed in Table I. A probable reason is that some of the attributes are unreliable, thus the constructed graph may have large weights between non-similar classes, reducing the discriminativity of the combined classifier.

2) Ablation Experiment on Structurally Constrained Correlation Transfer: To further illustrate the effectiveness of our method, we test our methods with one-vs-others strategy using different values of R_n on AwA, CUB and SUN, as shown in Fig. 2. From the figure, we can observe that as the values of R_n increase, the classification accuracy tends to present an up and down trend, which shows that the local connectivity constraint over embeddings of phantom classes successfully improves the classification accuracy, proving that our assumption that each class embedding and its neighbors lie on a locally linear patch on the manifold is valid. Note that model [4] can be considered as a special case of our method when $R_n = R$, which demonstrates that by considering the structural relations in the semantic embedding space, we can obtain a more discriminative classifier for unseen classes.



(a) (b)

Fig. 2. Performance results with different neighbor phantom classes R_n with one-vs-others strategy (the horizontal axis corresponds to the proportion of R_n/R).

3) The Number of Neighbor Phantom Classes R_n: To explore whether the values of R_n are the same in different datasets, we conduct experiments under different constraints. The values of R_n in different datasets are shown in Table III. We can preliminarily conclude that R_n increases with the increase of the amount of unseen classes and basically remains unchanged with different loss functions. We infer that the values of R_n are related to the data distribution and are independent of the loss constraints. Note that in the experiment, we set $R = S$ and we will further explore the effects of different values of R to verify our conclusion.

V. Conclusion

We propose to solve ZSL problem by introducing the structurally constrained correlation transfer into the manifold based semantic embedding framework. To strengthen the structural constraint of embeddings of similar classes and reserve the global structure in the semantic embedding space, the proposed method explores to form and project the neighbor-preserving structure from the semantic space to the model space and predict classifiers for unseen data. Experiments on three benchmark datasets validate the efficiency of the proposed method.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61425011, Grant 61720106001, Grant 61529101, and in part by Shanghai High Technology Project under Grant 17511106603 and the Program of Shanghai Academic Research Leader under Grant 17XD1401900.

References

[1] X. Yongqin, S. Bernt, and A. Zeynep, "Zero-shot learning-the good, the bad and the ugly," *arXiv preprint arXiv:1703.04394*, 2017.

[2] L. C. H, N. Hannes, and H. Stefan, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, 2014.

[3] N. Mohammad, M. Tomas, B. Samy, S. Yoram, S. Jonathon, F. Andrea, C. G. S, and D. Jeffrey, "Zero-shot learning by convex combination of semantic embeddings," in *Proc. ICLR'14*, 2014.

[4] C. Soravit, C. Wei-Lun, G. Boqing, and S. Fei, "Synthesized classifiers for zero-shot learning," in *Proc. CVPR'16*, 2016, pp. 5327–5336.

[5] R. S. T and S. L. K, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[6] A. Zeynep, R. Scott, W. Daniel, L. Honglak, and S. Bernt, "Evaluation of output embeddings for fine-grained image classification," in *Proc. CVPR'15*, 2015, pp. 2927–2936.

[7] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Proc. NIPS'03*, 2003, pp. 857–864.

[8] C. Koby and S. Yoram, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of machine learning research*, vol. 2, no. Dec, pp. 265–292, 2001.

[9] A. Zeynep, P. Florent, H. Zaid, and S. Cordelia, "Label-embedding for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1425–1438, 2016.

[10] M. Thomas, G. Efstratios, and S. C. GM, "Costa: Co-occurrence statistics for zero-shot classification," in *Proc. CVPR'14*, 2014, pp. 2441–2448.

[11] W. Catherine, B. Steve, W. Peter, P. Pietro, and B. Serge, "The caltech-ucsd birds-200-2011 dataset," 2011.

[12] P. Genevieve and H. James, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *Proc. CVPR'12*, 2012, pp. 2751–2758.

[13] K. Alex, S. Ilya, and H. G. E, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS'12*, 2012, pp. 1097–1105.

[14] S. Christian, L. Wei, J. Yangqing, S. Pierre, R. Scott, A. Dragomir, E. Dumitru, V. Vincent, and R. Andrew, "Going deeper with convolutions," in *Proc. CVPR'15*, 2015, pp. 1–9.

[15] J. Yangqing, S. Evan, D. Jeff, K. Sergey, L. Jonathan, G. Ross, G. Sergio, and D. Trevor, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM MM'14*, 2014, pp. 675–678.