LEARNING SIMILARITY-SPECIFIC DICTIONARY FOR ZERO-SHOT FINE-GRAINED RECOGNITION

Hong Chen[1], *Liujuan Cao*[1*], *Rongrong Ji*[2]

[1]Fujian Key Laboratory of Sensing and Computing for Smart City, Department of Computer Science, School of Information Science and Engineering, Xiamen University, China
[2]Fujian Key Laboratory of Sensing and Computing for Smart City, Department of Cognitive Science, School of Information Science and Engineering, Xiamen University, China
hongc@stu.xmu.edu.cn, caoliujuan@xmu.edu.cn, rrji@xmu.edu.cn

ABSTRACT

In this paper, we study the problem of zero-shot fine-grained recognition. It aims to distinguish unseen subordinate categories through some other seen categories within an entry-level category. We demonstrate the necessity to learn multiple latent dictionaries through joint training with specific set of instances, human-defined attributes and the class labels. A novel approach that is capable of 1) automatically assigning suitable dictionaries for each instance and 2) learning similarity-specific semantic representations for zero-shot fine-grained recognition is proposed. Experimental results on three benchmark datasets demonstrate that the proposed method achieves superior or comparable performance.

Index Terms— Image analysis, zero-shot learning, fine-grained recognition

1. INTRODUCTION

Zero-shot fine-grained recognition is an important issue that has many real-world applications. For example, it is well-known that object frequencies in natural images follow a long-tailed distribution [1], in which the uncommon objects do not occur frequently comparing to the common ones. For unseen or unfrequent categories, it needs heavy manpower labeling to collect and annotate sufficient training samples, especially for fine-grained tasks that need specialized domain knowledge [2]. The definition of this issue is to distinguish subordinate but unseen categories within an entry-level category, such as identifying new bird species or novel particular models of aircraft. To make it more suitable for zero-shot fine-grained recognition task, appropriate strategy is of great importance for zero-shot learning.

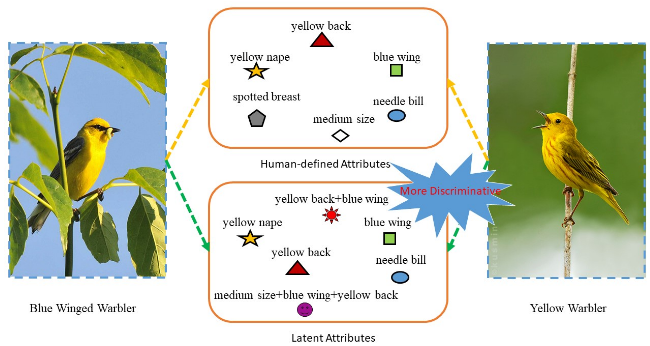

Fig. 1. Human-defined attributes vs. Latent attributes. The learned latent attributes would be more discriminative.

Zero-shot learning (ZSL) recognizes an object instance from a new category never seen before with the help of semantic cues, $e.g.$, human-defined attributes [3, 4, 5], text descriptions [6], word vectors [7]. The assumption for typical ZSL methods is that there exists a shared embedding space, in which a mapping function, $F(x, y; W) = \phi(x)^T W \psi(y)$, is defined to measure the compatibility between the image features $\phi(x)$ and the semantic representations $\psi(y)$ for both seen and unseen classes. W is the visual-semantic mapping matrix to be learned.

Zero-shot fine-grained recognition should satisfy two crucial criteria: 1) to be discriminative for different categories and 2) to inherit a good semantic space to efficiently classify novel categories. Most of the previous methods [4, 6, 8, 9] focus on the second criterion, are mainly driven by exploring a good alignment between the visual and semantic space, whilst the importance to learn discriminative representations is left unexploited.

Subsequently, in this paper we mainly focus on the following issues to be solved for zero-shot fine-grained recognition: firstly, the human-defined attributes, though seman-

*Corresponding author. This work is supported by the Nature Science Foundation of China (No.61772443, No.U1705262, and No.61572410), National Key R&D Program (No.2017YFC0113000, and No.2016YFB1001503), Post Doctoral Innovative Talent Support Program under Grant BX201600094, China Post-Doctoral Science Foundation under Grant 2017M612134, Scientific Research Project of National Language Committee of China (Grant No. YB135-49), and Nature Science Foundation of Fujian Province, China (No. 2017J01125 and No. 2018J01106).

tically descriptive, are not exhaustive and discriminative enough. For example, as shown in Fig. 1, in Caltech-USCD-Birds-200-2011 fine-grained retrieval dataset [2], each bird class is described by human-defined attributes such as *'yellow back', 'blue wing', 'medium size', etc*. These annotations are shared in many categories thus are desirable for knowledge transfer between categories, especially from seen to unseen categories. While to distinguish similar features requires the annotations to be discriminative to make the prediction more reliable. Peng $et\ al.$ [8] proposed to learn the latent attributes that are complementary to the human-defined attributes and combinate these two attributes together for richer representation. And Jiang $et\ al.$ [10] proposed to learn latent attribute dictionary jointly with attribute space and similarity space, which thus has more capacity in separating image features.

Secondly, the state-of-the-art ZSL methods [10, 11, 12] use a global embedding function for all types of images. These methods, though improve zero-shot recognition accuracy to some extent, they are not particularly suitable for fine-grained recognition task that needs a more compatible model for each indistinguishable feature [13]. Xian $et\ al.$ [13] randomly assigned training instances to study multiple bilinear compatibility functions to capture latent discriminative features and obtained considerable performance.

The aforementioned works improve zero-shot recognition accuracy through either exploring latent discriminative attributes or learning multiple bilinear compatibility functions. Our method take advantage of both that makes it more discriminative and more suitable for zero-shot fine-grained recognition task. Our contributions are three-fold:

- A simple but effective stratery is designed to augment the human-defined attributes and to learn similarity-specific dictionaries.
- A dictionary assignment phase is proposed to assign each test example to appropriate latent dictionary. Each example is evaluated through suitable dictionaries.
- Extensive experimental evaluations on three benchmark datasets show the effectiveness of the proposed method.

The remainder of this paper is organized as follows: task definition is introduced in Sec.2. Detailed descriptions of the proposed approach and quantitative experiments are given in Sec.3-Sec.4. Finally, we conclude this paper in Sec.5.

2. TASK DEFINITION

For a zero-shot fine-grained recognition task, the training set, *i.e.*, the seen classes, is defined as $S = \{(x_i^s, y_i^s)\}_{i=1}^{n^s}$, where $x_i^s \in X^s$ is a d-dimensional column vector representating the i-th training image from C^s seen classes, and $y_i^s \in Y^s$ is the corresponding label. The test set, *i.e.*, the unseen classes, is defined as $U = \{(x_j^u, y_j^u)\}_{j=1}^{n^u}$, where $x_j^u \in X^u$ is a d-dimensional column vector representating the j-th test image from C^u unseen classes, and $y_j^u \in Y^u$ is the corresponding label. Typically, label sets of seen classes and unseen classes are disjointed, *i.e.*, $Y^s \bigcap Y^u = \varnothing$. Additionally, the human-defined attributes for both seen and unseen classes are denoted as $A^s = \{a_i^s\}_{i=1}^{C^s}$ and $A^u = \{a_j^u\}_{j=1}^{C^u}$, where a_i^s and a_j^u indicate the attribute vectors for the i-th seen class and the j-th unseen class, respectively. At the test stage, given a test instance x^u and the attribute annotations of the test classes A^u, the goal is to predict the correspond category label y^u for x^u.

3. METHODS

The proposed method for zero-shot fine-grained recognition is illustrated in Fig. 2. Note that the architecture contains multiple iterative processes of latent discriminative dictionary learning. For clarity, we illustrate the process of learning one latent dictionary as an example. In each process, the procedure consists of three different components, 1) the deep feature network (DF-Net) to extract image features, 2) the appropriate dictionary assignment (ADA) to assign each instance to suitable latent dictionaries and 3) a similarity-specific dictionary learning phase to build the embedding space where the visual and semantic information are associated.

3.1. The Deep Feature Network

Previous works in the field of object recognition have demonstrated the success of deep convolutional networks in feature extraction. Therefore, our framework starts with a convolutional networks responsible for extracting image features, which is termed as DF-Net. Two kinds of the widely used networks are considered, $i.e.$,VGG-19 [14] and GoogLeNet [15]. For VGG-19, the DF-Net starts from *conv1* to *fc7*. For GoogLeNet, it starts from *conv1* to *pool-5* . The feature $\phi(x)$ of input image extracted from DF-Net can be formulated as:

$$\phi(x) = W_{DF} \star x, \tag{1}$$

where W_{DF} represents the overall parameters of the DF-Net, and $\star$ denotes a set of operations of DF-Net.

3.2. Appropriate Dictionary Assignment

Learning a single dictionary for ZSL typically leads to the inconsistency between dictionary and different kinds of indistinguishable features, as demonstrated in [13]. Inspired by the common observations that visually similar images are spatially nearby in visual feature space, we focus on different indistinguishable features to benefit the process of dictionary learning and category classification. Therefore, a new phase, termed appropriate dictionary assignment (ADA), is designed for discriminative dictionary learning, which assigns each dictionary with a set of visually similar examples.

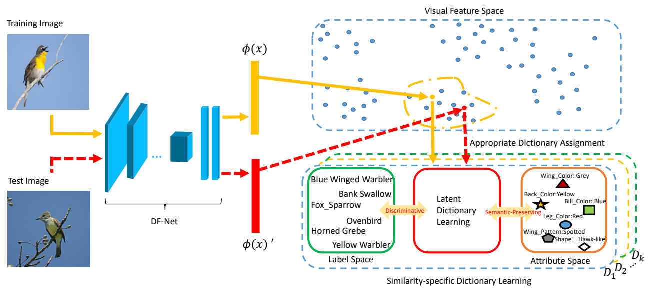

Fig. 2. Framework of the proposed method. The image representations extracted from deep convolutional network (DF-Net) are projected into visual feature space to assign to suitable dictionaries. The assignment process is based on the distances between each instance and centers of Gausian models. After this phase, each latent dictionary is jointly trained with human-defined attributes, label space and specific set of training instances. The learned dictionaries are discriminative, semantic-preserving and similarity-specific.

Because of the casual shape of distributions in visual feature space, Gaussian mixture model is adopted to modeling the distribution of examples. More specially, we hypothesize that each instance subjects to Gaussian distribution, then:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k), \ s.t. \ K \in N^+, \tag{2}$$

where π_k is the k-th mixing coefficient indicating the probility of instance x belonging to the distribution k. μ_k and Σ_k represent the learned mean and covariance for the k-th distribution, respectively. Eq. 2 is solved through EM algorithm to learn an optimal combination of $\{\pi_k, \mu_k\}$ and the corresponding Σ_k of each distribution. ADA takes the output of the last layer of DF-Net (*e.g.*, *pool-5* in GoogLeNet) and assigns it to suitable dictionaries based on the Euclidean distance between $\phi(x)$ and μ_k during the training and testing phases.

3.3. Similarity-specific Dictionary Learning

To be more adaptable for fine-grained recognition task, we introduce a joint dictionary learning phase to learn similarity-specific dictionaries with specific set of training instances, human-defined attributes and label space. The object function is formulated as follows:

$$\begin{aligned} \underset{D_k, Z_k, W_k, M_k}{\arg\min} \ & \|X_k - D_k Z_k\|_F^2 + \alpha \|Z_k - W_k A\|_F^2 \\ & + \beta \|Y_k - M_k Z_k\|_F^2 + \gamma \|D_k\|_F^2, \\ s.t. \ & \|d_i\|_2^2 \leq 1, \ \ \|w_i\|_2^2 \leq 1, \ \ \|m_i\|_2^2 \leq 1, \forall i, \end{aligned} \tag{3}$$

where $k = 1, 2, ..., K$ with $K \geqslant 2$ indexes over the latent dictionaries.

X_k is a set of training examples selected by ADA. Y_k and Z_k indicate the corresponding labels and the reconstruction coefficients for the k-th latent dictionary, respectively. D_k is the learned latent dictionary for the k-th feature distribution. W_k constructs the relationship between the latent attribute dictionary and human-defined attributes to make the learned latent dictionary semantic-preserving. By joint training of category labels, matrix M_k makes the learned dictionary discriminative that contribute to category classification phase. Parameter α and β control the strength of dictionary learning and semantic-preserving, respectively. γ is a positive regularisation parameter and is fixed as $\gamma = 1$ in this work.

3.4. Zero-Shot Prediction

In zero-shot fine-grained recognition, we verify the predicted class label given test image. Given an image x_j^u and the semantic representation A^u of C^u unseen classes, we obtain the feature vector through DF-Net as $\phi(x_j^u)$. The human-defined attributes $A^u = \{a_j^u\}_{j=1}^{C^u}$ are projected into the latent attribute space through matrix D^k. While $\phi(x_j^u)$ are projected into the same space by W_k. We allocate suitable learned dictionaries and combine all the information for label prediction by:

$$c_j = \arg\min_{c=1}^{C^u} \sum_{k=1}^{K} l_t \|W_k \phi(x_j^u) - D_k A^u\|_F^2 + \delta \|A^u\|_2^2, \tag{4}$$

where D_k is the k-th dictionary trained from the k-th set of training data evaluated by using the reconstruction error. δ is a regularization term that favours a smaller norm. In this study, we set $\delta = 1$ for simplicity.

The first setting, termed *Proposed-M1*, l_t is fixed as $l_t = 1$. Prediction results of different dictionaries are combined through naive ADD operation. While on the second setting, termed *Proposed-M2*, l_t is defined as $l_t = \frac{1}{t+1}$ to penalize dictionaries with lower rank, where t represents the t-th nearby dictionary in the embedding space. The distance is evaluated based on Euclidean distance between $\phi(x_j^u)$ and μ_k.

4. EXPERIMENTS

4.1. Experimental Settings

Quantitative experiments are conducted on three benchmark datasets, *i.e.*, Animal with Attributes (AwA) [3], Caltech-USCD-Birds-200-2011(CUB) [2], and SUN-A [16]. Details of the three datasets are shown in Table 1. We utilize the attributes provided by the original datasets.

AwA contains 30,475 images beloging to 50 animal classes, paired with a human provided 85-D attribute inventory and corresponding class-attribute associations. We follow the default split that has been provieded in [17]. CUB consists of 200 bird species with 11,788 images that serves

Table 1. Details of the three benchmark datasets. ('No.' represents for 'Number'.)

dataset	No. of attributes	No. of seen classes	No. of unseen classes	No. of train samples	No. of test samples
AwA [3]	85	40	10	24295	6180
CUB [2]	312	150	50	7057	2933
SUN-A [16]	102	707	10	14140	200

as a benchmark dataset for fine-grained recognition and retrieval. We use the same zero-shot split as [18]. The SUN-A dataset was introduced by Patterson and Hays in [16], which is a subset of the SUN Database [19]. It is a fine-grained dataset, which shows less variations across different classes.

In this paper, two parameters α and β are tuned using five-fold cross-validation. The size of latent dictionary is fixed as 600. K is determined empirically. Our experiments show that $K = 11$ can fully capture different kinds of indistinguishable features, distinguishing them clearly. We use the common evaluation metrics of ZSL, *i.e.*, the multi-class classification accuracy (MCA) to evaluate the models:

$$MCA = \frac{1}{|N|}\sum_{i=1}^{|N|} class_i, \tag{5}$$

where $class_i$ is the prediction accuracy of i-th unseen class. $|N|$ corresponds the total number of unseen classes.

4.2. Experimental Results

The comparison results are shown in Table 2. Among all the comparisons, DAP [17], ESZSL [9] and SSE [20] study a fixed mapping between semantic space and visual space, while LAD [10] jointly learns a latent dictionary with semantic space, visual space and category labels. The increases of MCA profit from the application of joint dictionary learning with category labels. Compared with the previous methods, LatEM [13] presents the first work to employ multiple projections to implicitly capture different visual characteristics of objects. The improvement of MCA demonstrate the advantage of using multiple dictionaries. Our framework take advantage of the two streams and propose to learn multiple similarity-specific dictionaries, which shows significantly better performance comparing to the methods using either of them.

Compared with *Proposed-M1*, the only difference between *Proposed-M1* and *Proposed-M2* is that *Proposed-M2* penalizes the irrelavent dictionaries. As is shown in Table 2, the performance of *Proposed-M2* consistently outperforms *Proposed-M1* on three datasets, which demonstrates that treating multiple similarity-specific dictionaries with different reliabilities that based on their interdependency benefits the recognition phase.

Several recently developed state-of-the-art approaches in the literature are selected for further comparison. As shown in Table 2, the proposed methods perform comparable or superior on all the datasets. In general, *Proposed-M2* based on VGG-19 and GoogLeNet achieve comparable accuracy on AwA dataset (82.7% vs. 82.4%). On CUB dataset, *Proposed-M2* based on VGG-19 achieves a MCA of 58.5%, which is higher than the best comparison LAD [10](56.6%) by 1.9%. Our model obtains more significant improvement and achieves 88.5% that outperforms all the comparison methods to the state-of-the-art method on SUN-A dataset. CUB and SUN-A are benchmark datasets for more challenging task, more specially, zero-shot fine-grained recognition. Our method consistently outperform all the baseline methods and achieve the best performance in zero-shot fine-grained recognition task.

Table 2. Zero-shot recognition accuracy $(\%)$ of the comparisons on three benchmark datasets. There are two kinds of features: VGG-19 features [V] and GoogLeNet features [G]. Note that some comparative approaches conduct experiments on other datasets or with other kinds of features, and we do not list those results. '-' indicates results not reported.

Method	AwA		CUB		SUN-A	
	V	G	V	G	V	G
DAP [17]	57.2	60.5	39.8	39.1	72.0	-
ESZSL [9]	75.3	59.6	-	44.0	82.1	82.1
SSE [20]	68.8	-	43.7	-	54.5	-
LatEM [13]	-	71.9	-	45.5	-	-
JLSE [21]	80.5	-	42.1	-	83.8	-
Long $et.al$ [12]	82.1	-	45.7	-	86.5	-
MFMR [22]	79.8	**76.6**	47.7	46.2	-	-
LAD [10]	82.4	-	56.6	-	85.0	-
Proposed-M1	80.0	67.1	56.8	**58.3**	81.5	84.5
Proposed-M2	**82.7**	76.1	**58.5**	**58.3**	**87.5**	**88.5**

5. CONCLUSION

In this work, we propose to pay attention to the more challenging recognition task, termed zero-shot fine-grained recognition. A novel framework that is more suitable for zero-shot fine-grained recognition has been proposed to study latent dictionaries that obtaining latent similarity-specific attributes for different types of visually indistinguishable features. We conduct comprehensive empirical analysis on three benchmark datasets and demonstrate the superiority of the proposed model.

6. REFERENCES

[1] Soravit Changpinyo, Wei Lun Chao, Boqing Gong, and Fei Sha, “Synthesized classifiers for zero-shot learning,” pp. 5327–5336, 2016.

[2] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie, “The caltech-ucsd birds-200-2011 dataset,” 2011.

[3] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling, “Learning to detect unseen object classes by between-class attribute transfer,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 951–958.

[4] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid, “Label-embedding for image classification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 7, pp. 1425–1438, 2016.

[5] Yan Li, Junge Zhang, Jianguo Zhang, and Kaiqi Huang, “Discriminative learning of latent features for zero-shot recognition,” *arXiv preprint arXiv:1803.06731*, 2018.

[6] Scott Reed, Zeynep Akata, Bernt Schiele, and Honglak Lee, “Learning deep representations of fine-grained visual descriptions,” pp. 49–58, 2016.

[7] Shay Deutsch, Soheil Kolouri, Kyungnam Kim, Yuri Owechko, and Stefano Soatto, “Zero shot learning via multi-scale manifold regularization,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5292–5299.

[8] Peixi Peng, Yonghong Tian, Tao Xiang, Yaowei Wang, Massimiliano Pontil, and Tiejun Huang, “Joint semantic and latent attribute modelling for cross-class transfer learning,” *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[9] Bernardino Romera-Paredes and Philip Torr, “An embarrassingly simple approach to zero-shot learning,” in *International Conference on Machine Learning*, 2015, pp. 2152–2161.

[10] Huajie Jiang, Ruiping Wang, Shiguang Shan, Yi Yang, and Xilin Chen, “Learning discriminative latent attributes for zero-shot classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4223–4232.

[11] Yashas Annadani and Soma Biswas, “Preserving semantic relations for zero-shot learning,” *arXiv preprint arXiv:1803.03049*, 2018.

[12] Yang Long, Li Liu, Ling Shao, Fumin Shen, Guiguang Ding, and Jungong Han, “From zero-shot learning to conventional supervised classification: Unseen visual data synthesis,” pp. 6165–6174, 2017.

[13] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele, “Latent embeddings for zero-shot classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 69–77.

[14] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, “Going deeper with convolutions,” pp. 1–9, 2014.

[16] J. Hays and G. Patterson, “Sun attribute database: Discovering, annotating, and recognizing scene attributes,” in *Computer Vision and Pattern Recognition*, 2012, pp. 2751–2758.

[17] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling, “Attribute-based classification for zero-shot visual object categorization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2014.

[18] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele, “Evaluation of output embeddings for fine-grained image classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2927–2936.

[19] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba, “Sun database: Large-scale scene recognition from abbey to zoo,” in *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 2010, pp. 3485–3492.

[20] Zeynep Akata, Honglak Lee, and Bernt Schiele, “Zero-shot learning with structured embeddings,” *CoRR*, vol. abs/1409.8403, 2014.

[21] Ziming Zhang and Venkatesh Saligrama, “Zero-shot learning via joint latent similarity embedding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 6034–6042.

[22] Xing Xu, Fumin Shen, Yang Yang, Dongxiang Zhang, Heng Tao Shen, and Jingkuan Song, “Matrix tri-factorization with manifold regularizations for zero-shot learning,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2007–2016.