Learning Autoencoder of Attribute Constraint for Zero-Shot Classification

Kun Wang[1], Songsong Wu[1], Guangwei Gao[2], Quan Zhou[3], Xiao-Yuan Jing[1]
1. School of Automation , NJUPT
2. Institute of Advanced Technology , NJUPT
3. Key Lab of Ministry of Education for Broad Band Communication and Sensor Network Technology , NJUPT
Nanjing, China
kunw0221@126.com, sswu@njupt.edu.cn

***Abstract*—The goal of zero-shot classification (ZSC) is to classify target classes precisely based on learning a semantic mapping from a feature space to a semantic knowledge space. However, the learned semantic mapping is only concerned with predicting source classes. Applying the semantic mapping to target classes directly will suffer from the semantic shift problem. In this paper, we propose a novel method called autoencoder of attribute constraint (AOAC) to settle this problem. In AOAC, we adopt the encoder-decoder paradigm to learn the semantic mapping. Additionally, we take the inaccurate attributes of source images into consideration and generate virtual data to solve it. The experimental results on two challenging datasets show that our proposed AOAC can resolve the semantic shift problem effectively and also improve the computational speed significantly.**

***Keywords*-zero-shot classification; semantic mapping; semantic shift; autoencoder of attribute constraint; computational speed**

I. Introduction

Learning a multiclass classification [10, 19, 20] model usually requires abundant manually labeled data. However, it is expensive and unrealistic to provide target classes with a large number of training images. The researchers once tried to adopt few-shot learning [16] to tackle this problem and they have achieved some results. Recently, some scholars have proposed the most extreme zero-shot classification [2, 4, 6, 7, 11, 16, 18, 19] to deal with this problem and have made a significant breakthrough in this respect.

The purpose of ZSC is to build a suitable model for target classes from the given source classes. Target classes and the given source classes must be disjoint in ZSC. The key to ZSC is the knowledge about how target classes are semantically [11, 13, 16] related to source classes. The general practice is to introduce a sharable semantic knowledge space. Such a space can be a semantic word vector [5] or a semantic attribute space [1, 2, 3]. The former is to learn a vector representation for each class. This word vector can be learned from large-scale textual database like Wikipedia in an unsupervised way, based on an independent natural language modeling task. Compared with conventional human supervision, it encoders richer semantic relationships between classes. In semantic attribute space, ZSC first learns a semantic mapping from a feature space to a attribute space using source classes at training stage. The learned mapping is then directly used to project the feature representation of a target image into the semantic attribute space at test stage. Finally, ZSC can get the class label of the target image by calculating the distances [15, 17] between the achieved attribute representation of the target image and the prototype representation of each target class. In this paper, we adopt the semantic attribute space as the sharable semantic knowledge space. Recently, Professor Xiang from Queen Mary University of London has proposed a novel method called Semantic Autoencoder (SAE) [12] based on the attribute learning. In SAE, it learns an encoder to project the visual feature representations of source images into the semantic attribute space with a projection matrix. Then it adopts a decoder to reconstruct the original feature representations as far as possible. It learns the best semantic mapping by minimizing the reconstruction error. This method has achieved compelling results. However, we can find that the authors assume each the attribute representation of each source image is replaced by its prototype in SAE. This practice will bring about one error. Although attribute representations of all images belonging to a same class are around its prototype closely, but the prototype can't correspond with their real attribute representations completely because of various factors [8,19].

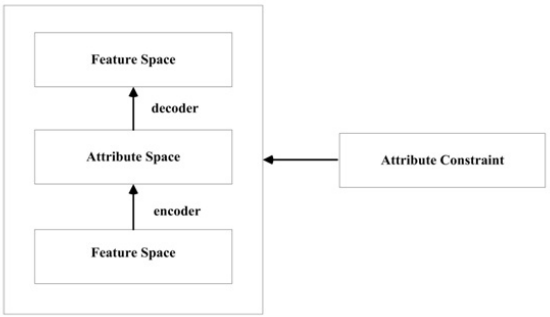

Figure 1: The proposed autoencoder of attribute constraint.

Actually, almost all methods based on the attribute learning suffer from the semantic shift problem [9, 14]. Those approaches mainly employ the learned semantic mapping to replace the real mapping for predicting target classes directly. Such a semantic mapping is only suitable for predicting source classes where relevant information of

2327-0985/17 $31.00 © 2017 IEEE
DOI 10.1109/ACPR.2017.129

target classes is missing. Direct use of the learned mapping affirmatively brings about the accuracy degradation in ZSC. This problem has attracted lots of attention from academia in recent years. In this paper, we will focus on the semantic shift problem to improve the ZSC accuracy.

In the work, we propose a novel autoencoder of attribute constraint mechanism to resolve the semantic shift problem, which is illustrated in Fig. 1. Specifically, we use the encoder to project the visual feature representations of source images into the semantic attribute space. We also adopt the decoder to project the achieved attribute representations back into the feature space and achieve the novel feature representations. We can achieve the best semantic mapping by minimizing the reconstruction error between the original feature representations and the new feature representations. Meanwhile, we can not ignore the fact that the given attribute representations of all source images can not correspond with their real attribute representations completely because of various factors. This situation will make our final model formulation inaccurate. We consider generating the virtual attribute representations of all source images to deal with this probelm in this paper. Under the proposed framework, the ZSC accuracy can be improved remarkably.

Our main contributions are: (1) A autoencoder of attribute constraint mechanism which can settle the semantic shift problem effectively is proposed for zero-shot classification. (2) Our proposed method is superior to the conventional ZSC methods in terms of the speed of calculation. Extensive experiments on two datasets demonstrate that our proposed AOAC can outperform significantly the other ZSC models.

II. Related Work

A. Attribute-based Zero-Shot Classification

The attribute space is shared among classes, so the knowledge can be transferred from source classes to target classes simply. In Direct/Indirect Attribute Prediction (DAP/IAP) [21], the attribute predictor can be learned by the given source classes and be used to obtain the attribute representation of a target image. Then ZSC achieves the class label by comparing the obtained attribute representation with the prototypes of all target classes. In Attribute Label-embedding Learning (ALE) [1, 13, 15], it learns an embedding space by maximizing the compatibility between images and labels. The above methods all get some promising results.

B. Relational Knowledge Transfer

Recently, some scholars have come up with a novel method called Relational Knowledge Transfer (RKT) [11] to deal with the semantic shift problem. In RKT, it extracts the semantic correlation between source classes and target classes in the attribute space on sparse coding theory. Then it transfers the semantic correlation to generate the virtual data of target classes in the feature space. The semantic mapping for target classes can be learned from only these generated data instead of source classes, which differs from our model. We also adopt the idea of generating the virtual data to tackle the inaccurate attribute representations of source images in our work.

C. Semantic Autoencoder

The semantic autoenoder (SAE) [12, 22, 23] has been an emerging topic in ZSC. A simple semantic autoencoder consists of a encoder, a decoder and one hidden layer which has a semantic meaning. The encoder aims to project the input data into the hidden layer with a projection matrix and the decoder is used to reconstruct the original input data as far as possible with another projection matrix. It can get the best semantic mapping by minimizing the reconstruction error.

D. Transductive Zero-Shot Recognition

Recently, the transductive learning [16, 30] has been a popular method in ZSC. It learns a class model which obtains the labels directly from images instead of attributes. In shared model space (SMS) [16], it has three important characteristics. Firstly, it learns a class model which directly generates class labels using images instead of attributes. Secondly, it proposes a shared model space for classes so that the class model can be learned easily using prototypes of classes. Thirdly, it also learns a joint learning mechanism which takes both source classes and target classes into consideration.

III. Method

A. Problem Definition

We are first provided with the given c_s source classes with N_s well-labeled source images $S = \{(x_1^s, z_1^s), (x_2^s, z_2^s), ..., (x_{N_s}^s, z_{N_s}^s)\}$. $X_s = [x_1^s, x_2^s, ..., x_{N_s}^s] \in R^{d \times N_s}$ is the visual feature matrix, where d indicates the dimension of feature and N_s is the number of source images. $Z_s = [z_1^s, z_2^s, ..., z_{N_s}^s] \in R^{k \times N_s}$ represents the semantic attribute matrix of the given source classes, where we define k as the dimension of attribute. We can assume that $Y_s = [y_1^s, y_2^s, ..., y_{N_s}^s] \in R^{k \times N_s}$ indicates our virtual attribute matrix of the source classes. Besides, we have the prototype matrix of the given source classes $P^s = [p_1^s, p_2^s, ..., p_{c_s}^s] \in R^{k \times c_s}$. We are also provided with a set of c_t target classes with N_t target images $T = \{(x_1^t, y_1^t), (x_2^t, y_2^t), ..., (x_{N_t}^t, y_{N_t}^t)\}$, where $Y_t = [y_1^t, y_2^t, ..., y_{N_t}^t]$ is the unknown attribute matrix of target images. The prototype matrix of target classes $P^t = [p_1^t, p_2^t, ..., p_{c_t}^t]$ is given at training stage. Our main purpose is to learn a suitable mapping which can calculate Y_t accurately from the given data.

B. Model

We adopt the semantic autoencoder paradigm in our work and presuppose the hidden layer is the semantic attribute space. We first project X_s into Y_s with a projection matrix $Q \in R^{k \times d}$. Then we project Y_s back into the feature space by another projection matrix $Q^* \in R^{d \times k}$ and achieve a novel visual feature matrix $X^{'} \in R^{d \times N_s}$.

We wish that $X^{'}$ is as similar as possible to X_s, so we write the objective as:

$$\min_{Q,Q^*} \|X_s - Q^* Q X_s\|_F^2 \;\; s.t. Q X_s = Y_s \tag{1}$$

We have two matrices to estimate in the objective and we can assume $Q^* = Q^T$ to simplify the above objective. Our objective thus becomes:

$$\min_{Q} \|X_s - Q^T Q X_s\|_F^2 \;\; s.t. Q X_s = Y_s \tag{2}$$

Now we only have one variable in (2). It will reduce the difficulty of simplification. In ZSC, we think that the real attribute representations of all images belonging to a same class are around its prototype representation. We use this principle to generate virtual attribute representations Y_s. We can rewrite (2) as:

$$\begin{aligned} &\min_{Q} \|X_s - Q^T Y_s\|_F^2 \\ &+\lambda_1 \|Q X_s - Y_s\|_F^2 \\ &+\lambda_2 \sum_{i=1}^{N_s} \sum_{j=1}^{c_s} m_{i,j} \|y_i^s - p_j^s\|_2^2 \end{aligned} \tag{3}$$

Where $m_{i,j} = 1$ if x_i^s belongs to the $j-th$ source class, otherwise $m_{i.j} = 0$, λ_1 and λ_2 are controlling coefficients in order to prevent the overfitting.

C. Optimisation

Obviously, (3) is only concerned with Q and Y_s. Thereby, we use the respective optimisation to resolve it. Specifically, we can turn the learned objective into two sub objectives:

$$\begin{aligned} &\min_{Y_s} \|X_s - Q^T Y_s\|_F^2 \\ &+\lambda_1 \|Q X_s - Y_s\|_F^2 \\ &+\lambda_2 \sum_{i=1}^{N_s} \sum_{j=1}^{c_s} m_{i,j} \|y_i^s - p_j^s\|_2^2 \end{aligned} \tag{4}$$

$$\min_{Q} \|X_s - Q^T Y_s\|_F^2 + \lambda_1 \|Q X_s - Y_s\|_F^2 \tag{5}$$

To optimise (4), we take a derivative of it and set it zero. We can get the final expression of Y_s:

$$Y_s = (Q X_s + \lambda_1 Q X_s + \frac{\lambda_2}{2} R_s^T)(I_{N_s} + \lambda_1 I_{N_s} + \lambda_2 I_{N_s})^{-1} \tag{6}$$

Where I is the identity matrix and the unknown R_s is written as follows:

$$R_s = \begin{bmatrix} (\sum_{j=1}^{c_s} 2 m_{1,j} P_j^s)^T \\ ... \\ (\sum_{j=1}^{c_s} 2 m_{N_s,j} P_j^s)^T \end{bmatrix} \tag{7}$$

We can reorganize (5) by using the following trace properties:

$$Tr(X_s) = Tr(X_s^T), Tr(Q^T Y_s) = Tr(Y_s^T Q) \tag{8}$$

Algorithm 1 Zero-Shot Classification by AOAC

Input:
- visual feature matrix of source images X_s
- visual feature matrix of target images X_t
- virtual attribute matrix of source images Y_s
- prototype of source classes P^s
- prototype of target classes P^t
- parameter $\lambda_1, \lambda_2, \alpha$

Output:
- attribute matrix of all target images Y_t

1: Initialize
   Y_s by P^s
   $m_{i,j} = 1$ if $x_i^s \in c_j^s$, otherwise $m_{i.j} = 0$
2: While not converge do
3: Update Q by (10)
4: Update Y_s by (6)
5: check the convergence condition:
   $\|Q_k - Q_{k-1}\|_2^2 \leq \alpha$
6: end

Equation. (5) then becomes:

$$\min_{Q} \|X_s^T - Y_s^T Q\|_F^2 + \lambda_1 \|Q X_s - Y_s\|_F^2 \tag{9}$$

Similarly, we take a derivative of (9) and set it zero:

$$Y_s Y_s^T Q + \lambda_1 Q X_s X_s^T = Y_s X_s^T + \lambda_1 Y_s X_s^T \tag{10}$$

We use the following formulas to simplify (10):

$$A = Y_s Y_s^T, B = \lambda_1 X_s X_s^T, C = (1 + \lambda_1) Y_s X_s^T \tag{11}$$

We can get the final expression of Q:

$$AQ + QB = C \tag{12}$$

Equation. (12) is the famous Sylvester equation. In MATLAB, it can be solved by the following formula:

$$Q = sylvester(A, B, C) \tag{13}$$

We achieve the right expressions of Q and Y_s. In the end, we believe that if Q_k is very close to Q_{k-1}, the Q is the best semantic mapping, so we adopt the following formula of iteration to get the best Q_k:

$$\|Q_k - Q_{k-1}\|_2^2 \leq \alpha \tag{14}$$

Where value α is given at training stage. Our model algorithm is summarised in Algorithm 1.

D. Classification

We can learn the best Q using the given data. Then we perform the classification which differs from the general practice. In this paper, we choose to project the prototypes of target classes into the visual feature space by $x^* = Q^T P^t$, where x^* are the prototype projections of all target classes in the feature space. We can get the class label of the target image x_i^t by calculating the distances between x_i^t and the prototype projections x^* :

$$f(x_i^t) = \underset{j}{argmin} \;\; d(x_i^t, x_j^*) \tag{15}$$

Where $x_j^* \in x^*$ is the $j-th$ prototype projection. d is a cosine distance function which returns the class label of the target image x_i^t.

Figure 2: The examples of AwA dataset.

Figure 3: The CUB dataset containing 200 different kinds of fine-grained birds.

IV. Experiments

In order to verify the superiority of our proposed method, we conduct extensive experiments on two benchmark datasets.

A. Experimental Setup

1) Datasets: We conduct a set of experiments on two benchmark datasets: Animals with Attributes (AwA) and Caltech UCSD Birds (CUB). AwA dataset consists of 30475 images from 50 coarse-grained [24] animals, such as "bird" and "cow" (Fig. 2). Each animal has at least 92 images and contains 85 binary attributes. CUB dataset contains 200 different kinds of fine-grained [24] birds, with 11788 images in total (Fig. 3). In CUB dataset, each class contains 312 binary attributes.

2) Features: For the past few years, deep features have been proven very effective in pattern recognition. In our experiments, we also use deep features extracted from the GoogLeNet architecture [20] to represent every image in both AwA dataset and CUB dataset.

3) Settings: We choose 40 animals as source classes and the others as target classes in AwA dataset. In CUB dataset, 150 classes and 50 classes are adopted as source classes and target classes respectively. In AwA dataset, we choose the first 80 source images of each source class as the novel source images, so we have 3200 source images in total. In the final objective, we have three parameters to evaluate. We get the best values for them by the constant debugging:

$$\lambda_1 = 150000, \lambda_2 = 0.1, \alpha = 25 \tag{16}$$

4) Competitors: We compare our method with 4 popular zero-shot classification approaches: RZSL; RKT; SMS and SAE. All above methods are proposed to solve the semantic shift problem and they have reached some promising results. We compare our proposed AOAC with these approaches in the same experimental environment.

Table I: ZSC accuracy (%).

Method	AwA	CUB
RZSL [18]	65.6	31.4
RKT [11]	66.2	38.4
SMS [16]	74.8	43.6
SAE [12]	81.4	57.4
AOAC	82.8	58.3

Table II: The computaional speed (s).

Method	Training	Test
RKT[11]	49.71	6.65
RZSL[18]	71.13	15.85
SMS[12]	59.63	11.65
AOAC	13.78	0.12

B. Experimental Results

We can draw the following conclusions from Table I and Table II: (1) Our AOAC model can get the best accuracy in both AwA dataset and CUB dataset. It improves 1.4% and 0.9% in AwA dataset and CUB dataset respectively. (2) The experimental results of our method in AwA dataset are better than which in CUB dataset. This is because CUB dataset includes sufficient images which are very similar. It is difficult for us to distinguish them completely. (3) Table II shows that whether at training stage or at test stage, our proposed method is the best in terms of the speed of calculating.

C. Analysis and Discussion

Extensive experiments on two datasets demonstrate that our proposed method is very prominent. It can effectively settle the semantic shift problem. Our proposed model tries to reconstruct the real mapping for predicting target classes as far as possible by using the semantic autoencoder paradigm. Additionally, we also decide to generate the virtual data to tackle the inaccurate attribute representations of source images. All above measures can help explain why our approach is superior to others.

In the model formulation, Q^* is equal to Q^{-1} actually and we assume $Q^* = Q^T$ to simplify the model. This practice affirmatively brings about one error. We believe that if we can minimize this error, the experimental results will be better than before. Besides, how to distinguish the abundant similar images in CUB dataset effectively still puzzles us. Those problems will be encouraged in our future work.

V. Conclusion

In the paper, we propose an autoencoder of attribute constraint mechanism to settle the semantic shift problem for ZSC. Our approach contains the following characteristics. Firstly, we adopt the encoder to project the visual feature representations into the semantic attribute space and then use the decoder to reconstruct the original data. Secondly, we achieve the best semantic mapping by minimizing the reconstruction error. Thirdly, we generate virtual data to resolve the inaccurate attribute representations of source images. The experimental results show that our proposed method can effectively solve the semantic shift problem.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant nos. 61402238 and 61502245, the Postdoctoral Science Foundation of Jiangsu Province under Grant no. 1302054C, the NUPTSF under Grant no. NY212029.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 819-826, 2013.

[2] D. Jayaraman, and K. Grauman. Zero-shot recognition with unreliable attributes. In Advances in Neural Information Processing Systems, 3464-3472, 2014.

[3] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S. Chang. Designing category-level attributes for discriminative visual recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 771-778, 2013.

[4] B. Romera-Paredes, and P. H. S. Torr. An embarrassingly simple approach to zero-shot learning. In ICML, 2152-2161, 2015.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, 3111-3119, 2013.

[6] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In NIPS, 1410-1418, 2009.

[7] R. Socher, M. Ganjoo, C. D. Manning, and A. Y. Ng. Zero-shot learning through cross-model transfer. In NIPS, 935-943, 2013.

[8] Y. Guo, G. Ding, X. Jin, and J. Wang. Learning predictable and discriminative attributes for visual recognition. In Proceedings of Twenty-Ninth AAAI Conference on Artificial Intelligence, 3783-3789, 2015.

[9] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In NIPS, 137-144, 2006.

[10] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2927-2936, 2015.

[11] D. Wang, Y. Li, Y. Lin, and Y. Zhuang. Relational Knowledge Transfer for Zero-Shot Learning. In Proceedings of Thirtieth AAAI Conference on Artificial Intelligence, 2145-2151, 2016.

[12] E. Kodirov, T. Xiang, and S. Gong. Semantic Autoencoder for Zero-Shot Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[13] M. Bucher, S. Herbin, and F. Jurie. Improving semantic embedding consistency by metric learning for zero-shot classification. In European Conference on Computer Vision, 730-746, 2016.

[14] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In ICCV, 2452-2460, 2015.

[15] Z. Zhang, and V. Saligrama. Zero-shot learning via semantic similarity embedding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4166-4174, 2015.

[16] Y. Guo, G. Ding, X. Jin, and J. Wang. Transductive Zero-Shot Recognition via Shared Model Space Learning. In Proceedings of Thirtieth AAAI Conference on Artificial Intelligence, 3434-3500, 2016.

[17] Z. Zhang, and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6034-6042, 2016.

[18] J. Yu, and S. Wu. Robust Zero-Shot Learning with Source Attributes Noise. In Proceedings of the 2016 International Conference on Progress in Informatics and Computing, 205-209, 2016.

[19] L. Wang, and S. Wu. Learning Discriminative Instance Attribute for Zero-Shot Classification. In Proceedings of the 2016 International Conference on Progress in Informatics and Computing, 209-213, 2016.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, 1097-1105, 2012.

[21] C. Lampert, H. Nickisch, and S. Harmeling. Attribute based classification for zero-shot visual object categorization. In IEEE TPAMI, 453-465, 2014.

[22] M. Chen, W. EDU, and Z. E. Xu. Marginalized denoising autoencoders for domain adaptation. In ICML, 2014.

[23] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In Proceedings of the 28th international conference on machine learning, 833-840, 2011.

[24] S. Reed, Z. Akata, B. Schiele, and H. Lee. Learning deep representations of fine-grained visual descriptions. In CVPR, 2016.

[25] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In CVPR, 951-958, 2009.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1-9, 2015.

[27] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. In NIPS, 2003.

[28] Y. Shigeto, I. Suzuki, K. Hara, M. Shimbo, and Y. Matsumoto. Ridge regression, hubness, and zero-shot learning. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 135-151, 2015.

[29] P. Luo, X. Wang, and X. Tang. A deep sun-product architecture for robust facial attributes analysis. In ICCV, 2864-2871, 2013.

[30] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In ECCV, 584-599, 2014.