Latent factor model based zero-shot learning

Y X Q Zhang¹, Y D Peng², J D Li², C X Zhang², D R Yu³, R T Xia³, D B Liu¹ and B Yang²
¹College of Software Engineering, Jilin University, Changchun 130012, China
²College of Computer Science and Technology, Jilin University, Changchun 130012, China
³School of Artificial Intelligence, Jilin University, Changchun 130012, China

Email: 598345200@qq.com

Abstract. The key of zero-shot learning is to design an appropriate approach to capture the potential relevance between the image visual features and category-level semantic knowledge. There is a lot of data with matrix form in zero-shot recognition problem, such as image label matrix, visual feature matrix and semantic information matrix. Inspired by the latent factor model which tackles recommendation problems well avoid being influenced by sparse matrix, we regard an image as a user and analogize each dimension of the visual feature as an item. The key to our approach is that we model the image recognition as a special recommendation problem by aid of latent factor model. We evaluate our algorithm on three classic benchmark data sets for both conventional and generalized zero-shot setting, the classification results outperform significantly the state-of-art approaches with low calculation cost.

Keywords: zero-shot learning, latent factor model, image recognition

1. Introduction

Though you have never seen a camel before, you can still identify it by giving the knowledge that an animal looks like a horse with two humps on back, this is zero-shot learning in real life. In recent years, there has been an increasingly rapid advance in the field of zero-shot learning, a case in point is the task of image recognition. The goal of zero-shot recognition is to learn a classifier that generalizes well on unseen categories when trained by seen categories[1]. Most existing methods follow the paradigm that learns a generic mapping function between the semantic space and visual feature space[2,3,4]. Although these methods have achieved good performances on classification accuracy, how to capture better potential relevance between visual feature space and semantic feature space can still be optimized. There are many matrices in zero-shot recognition, i.e. seen-sample label matrix, image visual feature matrix and class semantic information matrix, we surprisingly find that the recognition process can be regarded as calculating the unseen-sample label matrix. Inspired by recommendation system, we creatively attempt to model the training process as a matrix factorization problem, and apply latent factor model to build relevance between semantic knowledge and visual feature, this kind of design makes the model training process free from the influence of sparse matrix.

Our contributions are as follows:(1) First, we propose a simple yet effective zero-shot recognition approach which recovers the image feature matrix by using image preference matrix and class contribution matrix in a latent factor model framework.(2) Besides, we choose graph structure to model the relationships of semantic information and deliver the information between unseen categories and
unseen categories by aid of graph neural network. (3) Extensive experiments are evaluated on three classic data sets in zero-shot recognition, the results demonstrate that our method brings effective improvements on recognition accuracy and achieves state-of-art results in this field.

2. Latent factor model

Latent factor model (LFM) is a kind of matrix factorization algorithm in the domain of recommendation systems, the main idea is using some latent factors to associate user interests with item features. In recommendation problem, the original user-rating matrix is usually sparse, which means that the users have not interacted with most products. Models can not achieve ideal results when there are lots of blanks in training samples, even for SVD[5] which is proved to be effective, the sparse matrix makes it difficult to learn eigenvalues and eigenvectors. Nevertheless, LFM aims to mine some latent factors and then decompose the original user-rating matrix into user-preference matrix and item-contribution matrix, furthermore, the product of these two matrices recovers the original matrix. As long as the user interacts with some of the products, the model can be trained without dense matrix.

Specifically, the matrix $R$, $P$ and $Q$ represent the user-item rating matrix, and user-class matrix and item-class matrix respectively in equation (1). The task is to find the most appropriate matrix $P$ and $Q$ to make the product $\tilde{R}$ of them be approximately equal to matrix $R$.

$$\tilde{R}_{m \times n} = P_{m \times k}^T \cdot Q_{k \times n} \approx R$$

Moreover, $P_{u,k}$ indicates how much interest the user $u$ has in latent factor $k$, $Q_{i,k}$ indicates how much contribution the item $i$ makes to the latent factor $k$. Then, $R_{u,i}$ (the user $u$’s interest in the item $i$) can be computed by following equation (2).

$$Interest(u, i) = R_{u,i} = \sum_{k=1}^{K} P_{u,k} \cdot Q_{i,k}$$

3. Problem formulation & Method

3.1. Problem formulation

In conventional zero-shot recognition, there are two sets of data: A train set $D_s$, where there are $N_s$ labeled samples within $S$ seen categories symbolized as $D_s = \{ (x_i, y_i) \}_{i=1}^{N_s}$, $x_i$ is the input image and $y_i \in Y_s = \{1, 2, ..., k\}$ is the corresponding class label. The other is a test set $D_u = \{ x_i \}_{i=1}^{N_u}$ where all the $N_u$ samples are from target unseen categories $Y_u = \{k+1, ..., K\}$, which is non-overlapping with the source categories $Y_s$. Additionally, we have attribute vectors $a_s \in A$ here as the semantic space and $s \in S = Y_s \cup Y_u$. The task aims at learning a visual-semantic mapping $f$ to predict the label $y_i \in Y_u$ for $D_u$. As for generalized zero-shot recognition, the test set contains data from both seen and unseen categories and all test samples will be classified into one of all categories $Y = Y_s \cup Y_u$, which is a more difficult task. Our approach is designed for an inductive zero-shot learning (ZSL) problem, which means that there is no image of unseen classes can be seen in the training phase.

3.2. Latent factor model based ZSL
As illustrated in figure 1, we model the zero-shot recognition as a recommendation problem by aid of latent factor model. We factorize the original image-visual feature matrix \( \mathbf{I} \) into two matrices: image-label matrix \( \mathbf{U} \) which indicates the category of each image and category-visual feature matrix \( \mathbf{V} \). Accordingly, we regard an image as a user in the recommendation problem and analogize each dimension in visual feature as an item, then similarly analogize \( \mathbf{I} \) as the user-rating matrix \( \mathbf{R} \), \( \mathbf{U} \) as the user-rating matrix \( \mathbf{P} \), \( \mathbf{V} \) as the user-rating matrix \( \mathbf{Q} \). With the intuition that the label information in matrix \( \mathbf{U} \) can be regarded as the latent preference factor of the given image, the matrix \( \mathbf{V} \) means how much each dimension in visual feature contributes to its corresponding category, it will be learned from category-level semantic knowledge. By computing the product using the image-label matrix \( \mathbf{U} \) and the category-visual feature matrix \( \mathbf{V} \) above, we could restore a new image-visual feature matrix which presents the visual feature proportion of each image.

As for image-visual feature matrix \( \mathbf{I} \), we use the mature convolution neural network \( f_v(\cdot) \) to get the corresponding visual feature representation of each image, which has been proved to be trustworthy in supervised image recognition. As for matrix \( \mathbf{V} \), we construct a semantic graph structure to build up connections with all classes, each node represent a class, and the edge between nodes depicts the between-class similarity. Then we incorporate a graph neural network \( g_s(\cdot) \) for projecting the attribute vectors \( \mathbf{A} \) from semantic space to visual space, and the output represents the category-level visual feature. To evaluate the effect of matrix factorization, we choose the mean squared error as the objective function which is commonly used in recommendation system, formulated as:

\[
L_s = \frac{1}{N_s \cdot |\mathcal{C}|} \sum_{i=1}^{N_s} \sum_{j=1}^{|\mathcal{C}|} (f_v(x_i) - u_i v_j)^2
\]

(3)

\( u_i \) represents the label of the image \( i \), \( v_j \) represents the semantic representation of class \( j \), their product predicts the visual feature of image \( i \). \( f_v(x_i) \) is the reliable visual embedding extracted from convolution neural network. \( d \) is the dimension of visual feature.

When preforming the test stage, the inference process can be performed in three stages: First, for a test image feature \( x_i \), obtain the visual feature embedding \( x_i^v \) using \( f_v(\cdot) \). Second, project the semantic features \( a_s \) of all classes (seen classes is for generalized ZSL) into the visual space \( y_s^v \) using the learned semantic network \( g_s(\cdot) \). Last, the result \( y_i^v \) can be predicted by computing the cosine similarity with all \( a_s^v \) and choosing the most matched class.

\[
y_i^v = \arg\min_{s \in \mathcal{Y}_s} \frac{x_i^v \cdot a_s^v}{\|x_i^v\|_2 \cdot \|a_s^v\|_2}
\]

(4)

**Figure 1.** The architecture of the proposed algorithm. “vf” refers to “visual feature”.

As illustrated in figure 1, we model the zero-shot recognition as a recommendation problem by aid of latent factor model. We factorize the original image-visual feature matrix \( \mathbf{I} \) into two matrices: image-label matrix \( \mathbf{U} \) which indicates the category of each image and category-visual feature matrix \( \mathbf{V} \). Accordingly, we regard an image as a user in the recommendation problem and analogize each dimension in visual feature as an item, then similarly analogize \( \mathbf{I} \) as the user-rating matrix \( \mathbf{R} \), \( \mathbf{U} \) as the user-rating matrix \( \mathbf{P} \), \( \mathbf{V} \) as the user-rating matrix \( \mathbf{Q} \). With the intuition that the label information in matrix \( \mathbf{U} \) can be regarded as the latent preference factor of the given image, the matrix \( \mathbf{V} \) means how much each dimension in visual feature contributes to its corresponding category, it will be learned from category-level semantic knowledge. By computing the product using the image-label matrix \( \mathbf{U} \) and the category-visual feature matrix \( \mathbf{V} \) above, we could restore a new image-visual feature matrix which presents the visual feature proportion of each image.

As for image-visual feature matrix \( \mathbf{I} \), we use the mature convolution neural network \( f_v(\cdot) \) to get the corresponding visual feature representation of each image, which has been proved to be trustworthy in supervised image recognition. As for matrix \( \mathbf{V} \), we construct a semantic graph structure to build up connections with all classes, each node represent a class, and the edge between nodes depicts the between-class similarity. Then we incorporate a graph neural network \( g_s(\cdot) \) for projecting the attribute vectors \( \mathbf{A} \) from semantic space to visual space, and the output represents the category-level visual feature. To evaluate the effect of matrix factorization, we choose the mean squared error as the objective function which is commonly used in recommendation system, formulated as:

\[
L_s = \frac{1}{N_s \cdot |\mathcal{C}|} \sum_{i=1}^{N_s} \sum_{j=1}^{|\mathcal{C}|} (f_v(x_i) - u_i v_j)^2
\]

(3)

\( u_i \) represents the label of the image \( i \), \( v_j \) represents the semantic representation of class \( j \), their product predicts the visual feature of image \( i \). \( f_v(x_i) \) is the reliable visual embedding extracted from convolution neural network. \( d \) is the dimension of visual feature.

When preforming the test stage, the inference process can be performed in three stages: First, for a test image feature \( x_i \), obtain the visual feature embedding \( x_i^v \) using \( f_v(\cdot) \). Second, project the semantic features \( a_s \) of all classes (seen classes is for generalized ZSL) into the visual space \( y_s^v \) using the learned semantic network \( g_s(\cdot) \). Last, the result \( y_i^v \) can be predicted by computing the cosine similarity with all \( a_s^v \) and choosing the most matched class.

\[
y_i^v = \arg\min_{s \in \mathcal{Y}_s} \frac{x_i^v \cdot a_s^v}{\|x_i^v\|_2 \cdot \|a_s^v\|_2}
\]

(4)
4. Experiment

To verify the effectiveness of the proposed framework (LFM-ZSL), extensive experiments were conducted on conventional ZSL and generalized ZSL. We demonstrate on three widely used ZSL data sets: AwA2 [2], CUB [6], and SUN [7], which are under the new GBU setting (which removes the overlapped classes between unseen classes and classes in ImageNet) for the train/test splits. We adopt the pretrained ResNet50 model to extract visual features for images, each image will be represented as a 2048-dimension vector. As for semantic network, we choose a 2-layers graph convolution network as the embedding tool. We follow the experiment setting and evaluation protocol as in [8] for a fair comparison.

4.1. Conventional zero-shot recognition result

This section reports the comparison results on the conventional zero-shot recognition in terms of top-1 accuracy. As can be seen in table 1, we witness that our method performs well on both coarse-grained data set AwA2 and fine-grained data set CUB, SUN. This improvement shows that our method learns effective correlation between visual features and semantic features.

Table 1. Comparative results under the conventional ZSL setting(%).

| Method  | AwA2  | CUB  | SUN  |
|---------|-------|------|------|
| DAP[2]  | 46.1  | 40.0 | 39.9 |
| CONSE[3]| 44.5  | 34.3 | 38.8 |
| SSE[4]  | 61.0  | 43.9 | 51.5 |
| DeVISE[9]| 59.7  | 52.0 | 56.5 |
| SJE[10]| 61.9  | 53.9 | 53.7 |
| ALE[11]| 62.5  | 54.9 | 58.1 |
| SYNC[12] | 46.6  | 55.6 | 56.3 |
| PSR[13]| 56.0  | 63.8 | 61.4 |
| LFM-ZSL| 67.8  | 64.8 | 68.2 |

4.2. Generalized zero-shot recognition result

We report our results of the more realistic generalized zero-shot setting in table 2. Though our method cannot always achieve the best classification result on seen class (s), nevertheless, it still gains leading results on unseen class(u) and harmonic result(h), comparing to other methods. More specifically, taking CUB as an example, the harmonic mean accuracy increases by 10.3% effectively.

Table 2. Comparative results under the generalized ZSL setting(%).

| Method | AwA2 | CUB | SUN |
|--------|------|-----|-----|
| CONSE  | 0.5  | **90.6** | 1.0 | 1.6 | **72.2** | 3.1 | 6.8 | 39.9 | 11.6 |
| SSE    | 8.1  | 82.5 | 14.8 | 8.5 | 46.9 | 14.4 | 2.1 | 36.4 | 4.0 |
| DeVISE| 17.1 | 74.7 | 27.8 | 23.8 | 53.0 | 32.8 | 16.9 | 27.4 | 20.9 |
| SJE    | 8.0  | 73.9 | 14.4 | 23.5 | 59.2 | 33.6 | 14.7 | 30.5 | 19.8 |
| ALE    | 14.0 | 81.8 | 23.9 | 23.7 | 62.8 | 34.4 | 21.8 | 33.1 | 26.3 |
| SYNC   | 10.0 | 90.5 | 18.0 | 11.5 | 70.9 | 19.8 | 7.9 | 43.3 | 13.4 |
| PSR    | 20.7 | 73.8 | 32.3 | 24.6 | 54.3 | 33.9 | 20.8 | 37.2 | 26.7 |
| LFM-ZSL| **29.0** | 69.5 | **40.9** | **33.9** | 65.6 | **44.7** | **29.4** | **45.3** | **35.6** |
5. Conclusion
In this work, we show a novel point of view that zero-shot recognition task can be modeled as a special recommendation problem, and apply latent factor model to mine potential associations between semantic feature space and visual feature space. The recognition problem is solved by recovering the unseen-sample label matrix and free from the influence of sparse matrix. We evaluate the proposed approach on multiple benchmark data sets that are widely used in zero-shot recognition task, the results demonstrate that our framework brings much improvements on recognition accuracy. We believe that there is still a lot of potential to be explored in the field of combining zero-shot learning domain and recommendation system domain.

6. References
[1] Y. Fu, T. Xiang, Y. Jiang 2018 IEEE Signal Processing Magazine 35 112-125
[2] C. H. Lampert, H. Nickisch, and S. Harmeling 2014 IEEE Transactions on Pattern Analysis and Machine Intelligence 36 453–465
[3] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado and J. Dean 2014(Preprint arxiv 1312.5650)
[4] Z. Zhang and V. Saligrama.2015 ICCV (Santiago) pp 4166–74
[5] Koren Y 2008 KDD (Las Vegas) pp 426–434
[6] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie 2011
[7] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba 2010 CVPR (San Francisco) pp 3485-92
[8] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata 2019. IEEE transactions on pattern analysis and machine intelligence. 41 2251-65
[9] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolovet 2013 NIPS pp 2121–29
[10] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele 2015 CVPR (Boston) pp 2927–36
[11] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid 2016. IEEE transactions on pattern analysis and machine intelligence. 38 1425–38
[12] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. 2016 CVPR (Las Vegas) pp 5327–36
[13] Y. Annadani and S. Biswas 2018 CVPR (Salt Lake City) pp 7603–12