Method for improving zero-shot image classification

Xiangfeng Chen\(^1,2\), Wenbai Chen\(^1,2\), Chong Zhang\(^1\), Mengyao Lv\(^1\), Hu Han\(^3\)

\(^1\)School of Automation, Beijing Information Science & Technology University, Beijing, People's Republic of China
\(^2\)MOE Key Laboratory of Machine Perception, Beijing, People's Republic of China
\(^3\)College of Automation, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, People's Republic of China

E-mail: 1057133456@qq.com

Abstract: In order to improve the robustness of the similarity metric method of image classification, and reduce the complexity of the measure function, the Pearson correlation coefficient is introduced to improve the zero-shot image classification. Firstly, the mapping matrix from visual space to semantic space is learned by the training dataset, and the visual feature is aligned by it. Then in the semantic space, the similarity between features is calculated by the metric function, predicting the label of unseen class by the nearest neighbours. Experiments show that zero-shot image classification based on Pearson correlation coefficient is better than Euclidean distance and cosine similarity.

1 Introduction

The similarity metric [1] is a measure of the closeness between different individuals, and it is greatly significant to algorithms such as clustering and neighbourhood search. During the zero-shot classification [2], the prediction of unseen samples is realised by the similarity metric, which will directly affect the classification accuracy of the model. The zero-shot learning aims to recognise the class that has not been seen in the training dataset, which relied on the dependence on the labelled samples. The area has received more and more attention. Li [3] proposed an end-to-end deep embedded zero-shot learning framework using multimodal fusion technology. Due to the hub problem, the projection direction from semantic space to visual space perform better. For the domain discrepancy of cross-data sets, the literature [4] presented a zero-shot classification based on the dual visual-semantic mapping path, which optimised the migration of visual-semantic mapping using the semantic intra-manifold intrinsic. The zero-shot based on semantic autoencoder (SAE) proposed by Elyor Kodirov [5] alleviated domain drift. Similarity metric is the basis for many visual tasks, such as target tracking, classification, and registration in robotic applications [6]. Ciobanu [7] proposed a new similarity metric to evaluate the effect of image denoising. In addition, it was also used to estimate the healthy quality of skeleton [8]. Chen [9] proposed an online event detection and tracking method based on similarity learning for the limitations of semantic representation and the poor performance of event detection and tracking. Gai et al. [10] proposed deep nonlinear metric learning applied into face verification. The literature [11] introduced depth transfer metric learning solving the problem of cross-domain visual identification.

The above zero-shot image classification aims to improve the mapping from the visual space to the semantic space by adding constraint terms, thereby enhancing the recognition accuracy and alleviating the domain drift problem. However, the similarity metric of zero-shot learning has attracted less attentions. The model usually accepts Euclidean distance or cosine to calculate the similarity, which is a limitation for classification. Although the deep metric learning has generalisation ability, and need cost long time. Considering comprehensively, the Pearson correlation coefficient is introduced to realise zero-shot image classification in the paper, improving the accuracy and reducing the training complexity.

2 Improvement method of zero-shot image classification algorithm

The whole framework of zero-shot learning is shown in Fig. 1, which contains visual models, natural language models, and multimodal fusion models. Visual model is trained by the deep convolutional neural network, extracting the features of image; then the semantic word vector corresponding to the class labels is obtained through the natural language model; the features is projected into the semantic space using the multi-modal fusion; finally, the prediction of unseen class is realised by comparison of the similarity between the features.

2.1 Feature alignment based on SAE

Multimodal fusion aims to project feature data of different modalities into the same space. The linear mapping can easily cause cross-domain drift.

In feature learning, the autoencoder obtains a meaningful feature representation based on the reconstruction error for the input data, achieving feature alignment. The traditional autoencoder is unsupervised learning. For zero-shot image classification, add the constraints between visual and semantic features. The network structure of the SAE is shown as Fig. 2.

Suppose the input data \(X = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^{d \times N}\) including \(N\) \(d\)-dim feature vectors. The latent feature \(S = (s_1, s_2, \ldots, s_N) \in \mathbb{R}^{d \times N}\) is obtained by the mapping matrix \(W \in \mathbb{R}^{d \times k}\), which is projected into the original feature space by the \(W^* \in \mathbb{R}^{k \times d}\) producing the features \(X' = (x'_1, x'_2, \ldots, x'_N) \in \mathbb{R}^{d \times N}\). Therefore, the objective function is defined as

![Fig. 1 Whole framework of zero-shot learning](image-url)
\[
\min_w \| X - W^*W X \|_F^2 \quad \text{s.t. } WX = S
\]
(1)

Due to \( W^* = W^T \), the objective function is simplified as
\[
\min_w \| X - W^T S \|_F^2 \quad \text{s.t. } WX = S
\]
(2)

As the constraint of \( WX = S \) is too strong, the objective function can be written as
\[
\min_w \| X - W^T S \|_F^2 + \lambda \| WX - S \|_F^2
\]
(3)

\( \lambda \) is the weight coefficient, which determines the importance of both terms. The above formulation is a standard quadratic homogeneous equation and convex function with a global optimal solution. Using the trace properties of the matrix, the objective function is shown in
\[
SS^TW + \lambda WXX^T = SX^T + \lambda SX^T
\]
(4)

The objective function meets the definition of Sylvester's equation. Finally, the mapping matrix \( W \) is obtained by Bartels–Stewar algorithm.

2.2 Similarity metric analysis in semantic space

For the different features, the similarity metric function selected will directly have an effect on the results of classification, which only depended on the target task. Generally, suppose data point \( p = (p_1, p_2, \ldots, p_d) \) and \( q = (q_1, q_2, \ldots, q_d) \), and the distance function is defined as \( d(p, q) \), which needs meet some rule as the following:

- \( d(p, p) = 0 \),
- \( d(p, q) \geq 0 \),
- \( d(p, q) = d(q, p) \),
- Triangle rule.

The most common way to measure the similarity between numerical points is the Euclidean distance, which is defined as
\[
d(p, q) = \left( \sum_{i=1}^{2} (p_i - q_i)^2 \right)^{1/2}
\]
(5)

Euclidean distance is intuitive. In addition, the smaller the distance is, the greater the similarity is. Besides, it has nothing to do with the distribution of the data. If the amplitude of \( p_i \) direction is too large, it will enlarge the effect of \( p_i \) dimension, and it cannot accurately express the similarity between related data. Mahalanobis distance can eliminate the correlation of different dimensions and the nature of different scales. Suppose \( p \) and \( q \) obey the same distribution and the covariance matrix is \( \Sigma \), so Mahalanobis distance is defined as
\[
d(p, q) = \sqrt{(p - q)^T \Sigma^{-1} (p - q)}
\]
(6)

It can avoid interference of correlation variables. However, it is affected by the instability of the covariance. Consider using cosine similarity as a measure of similarity, expressed as in
\[
\text{sim}(p, q) = \frac{\Sigma_{ij} p_i q_j}{\sqrt{\Sigma_{ii} \Sigma_{jj}}} \left| p \right| \left| q \right|
\]
(7)

Cosine similarity is only related with the direction of vector. However, it is influenced by the translation of the vector. So the Pearson correlation coefficient is introduced to improve the similarity computation by centralise the original data. For all the elements of every vector, calculate an average value \( \mu_p \) and \( \mu_q \) and it can be defined as in

\[
\Phi(x_i) = \arg \max_j \text{sim}(s'_i, s_j), \quad i, j = 1, 2, \ldots, u
\]
(9)

\( s'_i \) is the visual feature in the semantic space of \( i \)-th sample, and \( S_j \) denotes the semantic ground truth of \( j \)-th class. The \( j \) is the label of a test sample with the max similarity.

3 Experimental results and analysis

3.1 Dataset

In this experiment, the animals with attributes 2 [12] are selected as the training dataset, which concludes 37,322 images and 50 classes, and the distribution of the number of samples for each class is shown in Fig. 3. In order to meet the requirement of zero-shot learning, it is necessary to ensure that the training set and the test set have no intersection. The training dataset consists of 40 class, and the rest ten classes are used as a test dataset. Each class has an 85-dimensional attribute [13] value, which is considered as the transferred knowledge between the seen and unseen class.

Currently, Due to the generalisation ability of deep volume and neural network, the visual features in the zero-shot learning is extracted by it. This experiment uses the 1024-dimensional visual features extracted by Googlenet [5].
3.2 Zero-shot image classification analysis

In the classification task, the accuracy of image classification can be enhanced using the meaningful and representative features. The features of the visual space are projected into the semantic space by a mapping matrix of SAE learning aligning with the features of the space, which can preserve the original visual information. Visualisation of the visual space and its corresponding features distribution in the semantic space is shown in Fig. 4 using the t-SNE [14].

It can be seen from Fig. 4 that SAE can not only align features, but also improve the distribution of visual features in the semantic space, shrinking the distribution interval of features of the same class, and contributing to predicting the labels of the unseen class through similarity measures.

The similarity metric is the last link in the zero-shot image classification, and it is also an important part. The choice of measurement method will directly affect the classification result. In the semantic space, the different metrics are selected in the experiment, whose results are shown in Table 1.

Table 1 compares different metric methods, including Euclidean distance, cosine similarity, and Pearson correlation coefficient. It can be concluded that the accuracy of the zero-sample classification model based on Euclidean distance is significantly

Table 1  ZSL classification accuracy and time-consuming

| Method                        | Accuracy, % | Time  |
|-------------------------------|-------------|-------|
| Euclidean distance            | 62.91       | 0.0867|
| cosine                        | 84.72       | 0.1218|
| Pearson correlation coefficient| 84.75       | 0.1163|

Fig. 3  Distribution of the number of samples for each class

Fig. 4  Distribution of visual features in two different spaces

(a) Visual feature distribution in vision space, (b) Visual feature distribution in semantic space

J. Eng., 2018, Vol. 2018 Iss. 16, pp. 1688-1691

This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)
lower than other methods, but it is the shortest time-consuming method. Compared with the cosine similarity with the Pearson correlation coefficient, the latter is better because the Pearson correlation coefficient has translation invariance and scale invariance, which is not affected by the amplitude.

The accuracy of the classification can reflect the relationship between the correct sample and the total number of test samples. It cannot intuitively reflect the samples of error recognition. Therefore, the confusion matrix is explored to evaluate the model based on Euclidean distance and Pearson correlation coefficient, which is shown in Fig. 5.

In the confusion matrix, the class IDs are [6, 13, 15, 18, 24, 25, 34, 39, 42, 48], and the correct classification of the sample is represented by the box on the main diagonal. The deeper it is, the more the number of correct classifications for the class sample is. From Fig. 4, the Pearson correlation coefficient effectively improves the classification accuracy.

4 Conclusion

The similarity metric method in zero-shot classification is studied in this paper. The Pearson correlation coefficient is introduced to measure the similarity of features, which has nothing to do with the amplitude of features. In addition, it can accurately express the similarity of data and is not affected by the correlation between variables. The similarity metric matrix between the data points can better reflect the distribution of data in the semantic space, and then predict the labels of unseen samples. The experimental results show that the zero-shot classification based on Pearson correlation coefficient is obviously superior to the Euclidean distance, and the accuracy is closed to cosine similarity, and it is less than its computational complexity. However, the Pearson correlation coefficient is a fixed measurement method. Next, the metric learning method of adaptive data is further studied.

5 References

[1] Demisse, G. G., Aouada, D., Ottersten, B.: ‘Similarity metric for curved shapes in Euclidean space’, 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, June 2016, pp. 5042–5050
[2] Lampert, C. H., Nickisch, H., Harmeling, S.: ‘Attribute-based classification for zero-shot visual object categorization’, IEEE Trans. Pattern Anal. Mach. Intell., 2014, 36, (3), pp. 453–465
[3] Zhang, L., Xiang, T., Gong, S.: ‘Learning a deep embedding model for zero-shot learning’, 2016, arXiv preprint arXiv:1611.05088
[4] Li, Y., Wang, D., Hu, H., et al.: ‘Zero-shot recognition using dual visual-semantic mapping paths’, 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, July 2017, pp. 5207–5215
[5] Kodirov, E., Xiang, T., Gong, S.: ‘Semantic autoencoder for zero-shot learning’, arXiv preprint arXiv:1704.08345, 2017
[6] Seo, J. H., Kwon, D. S.: ‘A skeletal similarity metric for quality evaluation of retinal vessel segmentation’, IEEE Trans. Med. Imaging, 2018, PP, (99), pp. 1–1
[7] Chen, G., Kong, Q., Mao, W.: ‘Online event detection and tracking in social media based on neural similarity metric learning’. IEEE Int. Conf. on Intelligence and Security Informatics, 2017, pp. 182–184
[8] Cai, X., Wang, C., Xiao, B., et al.: ‘Deep nonlinear metric learning with independent subspace analysis for face verification’. ACM Int. Conf. on Multimedia, 2012, pp. 749–752
[9] Hu, J., Lu, J., Tan, Y. P.: ‘Deep transfer metric learning’, IEEE Trans. Image Process., 2016, 25, (12), pp. 5576–5588
[10] Xiao, Y., Lampert, C. H., Schiele, B., et al.: ‘Zero-shot learning – a comprehensive evaluation of the good, the bad and the ugly’, 2017
[11] Han, H., Jain, A. K., Shan, S., et al.: ‘Heterogeneous face attribute estimation: a deep multi-task learning approach’, IEEE Trans. Pattern Anal. Mach. Intell., Early Access 2017, p. 1
[12] Van Der Maaten, L.: ‘Accelerating t-SNE using tree-based algorithms’, J. Mach. Learn. Res., 2014, 15, (1), pp. 3221–3245