

Chapter 9

Scene Classification Based on Regularized Auto-Encoder and SVM

Yi Li, Nan Li, Hongpeng Yin, Yi Chai and Xuguo Jiao

Abstract Scene classification aims at grouping images into semantic categories. In this article, a new scene classification method is proposed. It consists of regularized auto-encoder-based feature learning step and SVM-based classification step. In the first step, the regularized auto-encoder, imposed with the maximum scatter difference (MSD) criterion and sparse constraint, is trained to extract features of the source images. In the second step, a multi-class SVM classifier is employed to classify those features. To evaluate the proposed approach, experiments based on 8-category sport events (LF data set) are conducted. Results prove that the introduced approach significantly improves the performance of the current popular scene classification methods.

Keywords Scene classification · Feature learning · Regularized auto-encoder · MSD · SVM

9.1 Introduction

In the last decades, scene classification has been an active and important research topic in image understanding [1]. It manages to automatically label an image among several categories. Scene classification can be applied to a wide spread application,

Y. Li · N. Li · H. Yin (✉) · Y. Chai · X. Jiao
School of Automation, Chongqing University, Chongqing 400044, China
e-mail: yinhongpeng@gmail.com

H. Yin
Key Laboratory of Dependable Service Computing in Cyber Physical Society,
Ministry of Education, Chongqing 400030, China

Y. Chai
Key Laboratory of Power Transmission Equipment and System Security,
Chongqing 400044, China

such as image indexing, object recognition, and intelligent robot navigation. Many approaches have been successfully adopted in scene classification [2]. However, in view of the variability between different classes and the similarity within one class, scene classification still remains a challenging issue.

In this paper, a novel scene classification method based on regularized auto-encoder and SVM is presented. It is simple but effective. The regularized auto-encoder is imposed with the Maximum Scatter Difference (MSD) criterion [3] and the sparsity constraint [4]. Scene classification based on this novel method mainly contains two stages, feature learning and classifier designing. The feature learning stage is based on the regularized auto-encoder. Features learned in this step are sufficient and appropriate to represent the source images. The classifier designing is based on a multi-class SVM, which is an acknowledged classifier that can achieve good performance in classification tasks. Experiments based on the LF data set demonstrate that the introduced method outperforms traditional methods in scene classification. There are two main contributions within our proposed scene classification method.

- (1) The auto-encoder with the sparsity constraint automatically extracts features from the source images. The learned features are sufficient and appropriate to describe the scene images. It is independent of prior knowledge, which can largely reduce the calculation cost.
- (2) The MSD criterion considers the similarity between different scenes and the dissimilarity within a scene. In the MSD work, images belonging to different categories can be easily classified by finding the best projection direction.

The next sections of the paper are organized as follows. The details of the novel scene classification approach are elaborated in Sect. 9.2. Experiment based on the LF data set is performed and result analysis is illustrated in Sect. 9.3. Conclusions and further research are summarized in Sect. 9.4.

9.2 The Details of the Proposed Scene Classification Approach

In this part, the novel scene classification approach is described in detail. This new proposed method contains steps of feature learning and classification. In the feature learning step, sufficient and appropriate features of source images are learned by the regularized auto-encoder imposed with the sparsity constraint and the MSD criterion. In the classification step, a multi-class SVM is adopted. Particularly, several categories of scene images are classified with the trained SVM classifiers. SVMs are trained by the 1-vs-1 strategy: constructing one SVM for each pair of the classes.

9.2.1 Feature Learning

Training images are disposed into small image patches. These patches suffer from measures of normalization and whitening. Then a regularized auto-encoder is trained by these pre-processed patches. Considering the big similarity between different categories and the small dissimilarity within a category, the proposed model is applied to extract sufficient and appropriate feature from the scene images.

The regularized auto-encoder model described in this paper is a 3-layer neural network, with an input layer and output layer of equal dimension, and a single hidden layer with k nodes. In particular, in response to the input patches set $X = \{x_1, \dots, x_m\}, x_i \in R^N$, the feature mapping $a(x)$ of the hidden layer with nodes, i.e. the encoding function is defined by Eq. (9.1).

$$a(x) = g(W_1 + b_1) \quad (9.1)$$

where $g(z) = 1/(1 + \exp(-z))$ is the non-linear sigmoid activation function applied component-wise to the vector z . And $a(x) \in R^K, W_1 \in R^{K \times N}, b_1 \in R^K$ are the output values, weights and bias of the hidden layer, respectively. W_1 is the bases learned from input patches. In order to make the input patches set less redundant, pre-processing is applied to X . In particular, after X is normalized by subtracting the mean and dividing by the standard deviation of its elements, the entire patches set may be whitened [5]. After pre-processing, the linear activation function of the output layer, i.e. the decoding function is

$$\tilde{x} = W_2 a(x) + b_2 \quad (9.2)$$

where $\tilde{x} \in R^N$ is the output value of the output layer. $W_2 \in R^{N \times K}$ and $b_2 \in R^N$ are the weight matrix and bias vector of the third layer, respectively. Thus the training of the regular auto-encoder model turns out to be the following optimization problem.

$$J_{re} = 0.5 \sum_{i=1}^m \|x_i - \tilde{x}_i\|^2 \quad (9.3)$$

Minimizing the above squared reconstruction error function with the back propagation algorithm, the weight matrices W_1, W_2 and bias b_1, b_2 are adapted. In order to overcome the over-fitting proble, a weight decay term, Eq. (9.4) is added to the cost function.

$$J_{wd} = 0.5\lambda(\|W_1\|_F^2 + \|W_2\|_F^2) \quad (9.4)$$

where $\|\cdot\|_F$ is the F -norm of W_1 and W_2 . Besides, λ represents the weight decay parameter.

The regular auto-encoder talked above relied on the number of hidden units being small. But there is a larger number of hidden units than the input pixels, it fails to discover interesting structure of the input. Particularly, when applied in multi-classification tasks, the similarity between different scenes and the dissimilarity within a scene largely reduce the classification accuracy. To handle the presented problems, we impose the sparsity constraint and the MSD criterion respectively.

(1) The sparsity constraint

The regular auto-encoder can easily obtain sufficient features of the input data. However, when the number of hidden units is large, the performance of regular auto-encoder is affected by the large computational expense. To discover interesting structure of the input data in the situation of lots of hidden units, a sparsity constraint is imposed on the hidden units. This novel auto-encoder model is known as sparse auto-encoder (SAE) [4]. Within the SAE framework, a neuron is treated as being “active” sparsity is imposed by restricting the average activation of the hidden units to a desired constant ρ . Specifically, this is achieved by adding a penalty term with the form $\sum_{i=1}^k KL(\rho||\hat{\rho}_i)$ to the cost function, where KL is the KL -divergence between ρ and $\hat{\rho}_i$. And is the sparsity parameter, $\hat{\rho}_i = (1/m) \sum_{i=1}^m a_j(x_i)$ is the average activation of hidden node j (averaged over the training set), and β controls the weights of the sparsity penalty term.

Thus the training of the regularized auto-encoder model turns out to be the following optimization problem:

$$J_{SAE} = \min(0.5 \sum_{i=1}^m \|x_i - \tilde{x}_i\|^2 + 0.5\lambda(\|W_1\|_F^2 + \|W_2\|_F^2) + \beta \sum_{j=1}^K KL(\rho||\hat{\rho}_j)) \quad (9.5)$$

(2) The MSD criterion

The Maximum Scatter Difference (MSD) norm [6] is a normalization of fisher discriminant criterion. It tries to seek a best projection direction, which can easily divide the categories of samples.

Assuming there are N pattern classes leave to be recognized, l_1, l_2, \dots, l_N , the intra-class scatter matrix S_b and inter-class scatter matrix S_w are defined as:

$$S_b = \frac{1}{C} \sum_{i=1}^N C_i (t_i - t_{mean})(t_i - t_{mean})^T \quad (9.6)$$

$$S_w = \frac{1}{C} \sum_{i=1}^N \sum_{j=1}^{C_i} (x_i^j - t_i)(x_i^j - m_i)^T \quad (9.7)$$

where C represents the training samples number, similarly C_i is the training samples number in class i , in which the j th training sample is denoted by x_i^j . Furthermore, t_{mean} and t_i are the mean vector of all training samples and training samples in class i respectively. Different categories are divided remotely with larger S_b value. Meanwhile, images from the same category get closer with smaller S_w value.

Reviewing the classic Fisher discriminant analysis [7], samples can be easily separated when the ratio of S_b and S_w , or their difference value gets maximal value. Combining with the above contents, we adopt the following format of MSD criterion:

$$J_{MSD} = w^T S_b w - \zeta \cdot w^T S_w w = w^T (S_b - \zeta \cdot S_w) \quad (9.8)$$

The two items $w^T S_w w$ and $w^T S_b w$ are balanced by the nonnegative constant ζ .

Based on the concept of Rayleigh quotient and its extreme property, the optimal solution of quotion (9.6), referring to the eigenvectors w_1, w_2, \dots, w_k , can be acquired by the characteristic equation $(S_b - \zeta \cdot S_w)w_j = \lambda_j w_j$, in which the first largest eigenvalues meet the requirements $\lambda_1 \geq \lambda_2 \geq \dots \lambda_k$.

Comparing with Fisher discriminant analysis method, the calculation of $S_w^{-1} S_b$ is replaced with $S_b - \zeta \cdot S_w$ within the framework of MSD criterion. Thus it will be fine weather S_w is a singular matrix or not. This makes computing becomes more effective.

Given the basic idea of auto-encoder, imposed with the sparsity constraint, the MSD criterion is imposed by taking the weight W_1 , which contacts the input layer with the hidden layer, as the projection matrix. Thus the proposed model turns to be the following optimization problem

$$J(w) = J_{SAE} + J_{MSD} \quad (9.9)$$

where J_{SAE} and J_{MSD} are described there-in-before. By a number of iteration steps, a balanced value of the weight and projection direction is obtained. With the proposed feature learning model, sufficient and appropriate features from training and testing images can be extracted.

9.2.2 Classifier Designing

In this part, a multi-class support vector machine (SVM) classifier is designed for scene classification. It is a kind of supervised machine learning. SVM is often used to learn high-level concepts from low-level image features. In this section, the one-against-one strategy is applied to train $l(l-1)/2$ non-linear SVMs, where l is the number of the scene categories.

Given the training data $v_i \in R^n$, $i=1, \dots, s$ in two classes, with labels $y_i \in \{-1, 1\}$. Each SVM solves the following constraint convex optimization problem

$$\min_{\psi, b, \xi} (0.5 \|\psi\|^2 + C \sum_{i=1}^S \xi_i), \text{ s.t. } y_i(\psi^T \phi(v_i) + b) \geq 1 - \xi_i, \xi_i > 0 \quad (9.10)$$

where $C \sum_{i=1}^S \xi_i$ is the regulation term for the non-linearly separable data set, $(\psi^T \phi(v) + b)$ is the hyper-plane. There are two main parameters that play an important role in SVM classification, C and γ . Parameter C represents the cost of penalty, which has great influence on the classification outcome. The selection of γ can affect the partitioning outcome in the feature space. Parameters C and γ make SVM achieve significant performance in classification tasks.

9.3 Experiment and Results Analysis

In this section, experiment on LF data set is conducted to evaluate the consequence of the introduced method. Experimental results show that better performance over current approaches is achieved in scene classification task (Fig. 9.1).

Experiment on LF data set

All these RGB images are firstly transferred into gray ones. Then feature learning is conducted with the proposed regularized auto-encoder model, convolution and mean pooling method. The obtained features are fed to multi-class SVM classifier to classify features. Figure 9.2 displays the confusion matrix of the sport data set.

The average performance obtained in 10 independent experiments is 91.56 %.

As shown in columns of Fig. 9.2, wrong classifications often occur in “badminton” and “bocce”, which demonstrates that in the feature space of the proposed



Fig. 9.1 There are 1579 RGB sport images included in LF data set: badminton (200), bocce (137), croquet (236), polo (182), rockclimbing (194), rowing (250), sailing (190), snow boarding (190). Each image has the size of 256×256 pixels

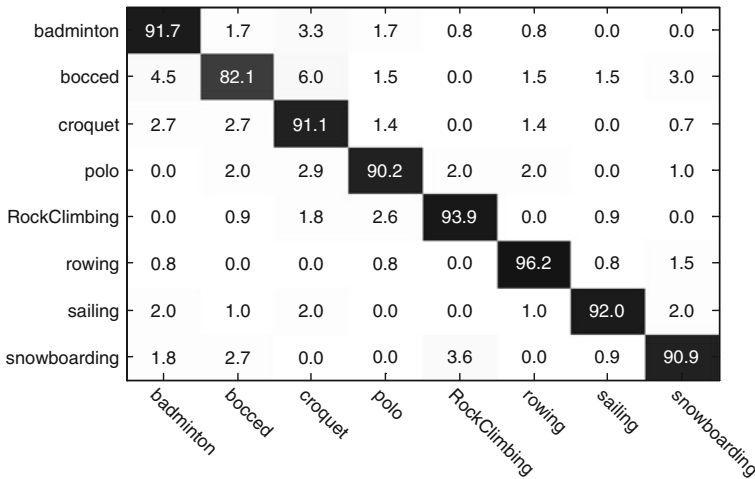


Fig. 9.2 The classification accuracy on LF data set

method, there are some similarities between these two categories with the other six categories.

Effect contrast with other current methods

In this section, experimental results of our method are compared with the previous studies on LF data set. The relevant researches and outcomes are listed in Table 9.1.

Seen from Table 9.1, our proposed method achieves better performance than the best performance Zhou et al. can achieve by far. Within the work of Zhou et al., a multi-resolution bag-of-features model is utilized. And the corresponding performance is an 85.1 % correct. Also can be seen from the outcome table, Li et al. get 73.4 % correct. This is realized by integrating scenes and object categorizations. It is mentionable that before long, Li et al. improve this performance by 4.3 % percentage points. In this method, objects of scene images are regarded as attributes, which makes recognition easier in scene classification.

Table 9.1 Effect contrast with previous studies

No	Methods	Accuracy (%)
1	Li et al. [8]	73.4
2	Wu et al. [9]	84.2
3	Li et al. [10]	77.9
4	Nakayama et al. [11]	84.4
5	Li et al. [12]	76.3
6	Zhou et al. [13]	85.1
7	Gao et al. [14]	84.9
8	Our approach	91.56

To analyze the improvement of our method, traditional approaches ignore the similarity between different categories and the differentiation between pictures within a category. Put another way, there may be people in bocced scene and bocced scene. Moreover, pictures in the badminton scene may contain people while other may not. Classification performance may be influenced to a certain extent by the object consistency problem. Fortunately, it can be solved by the MSD criterion adopted in our work. Significant performance has been proved by the experimental results based on the LF data set.

9.4 Conclusions

In this paper, a simple but effective scene classification method is proposed. It is based on a regularized auto-encoder model. Comparing with previous approaches, the proposed method achieves better performance due to three major improvements. First, the regular auto-encoder is imposed with the sparsity constraint, which is known as SAE model criterion. It automatically learns image features without relying on prior knowledge. Second, the MSD criterion is added into the SAE model, considering the similarity between different scenes and the dissimilarity within a scene. It offers more discriminative information of the input images. Thus, sufficient and appropriate features with more discriminative information are automatically extracted without prior information. Third, a multi-class SVM classifier is designed with the usually used optimization algorithm. It improves the classification accuracy efficiently. Results on LF data set indicate that this scene classification method gains better classification accuracy than other current methods.

References

1. Serrano-Talamantes JF, Aviles-Cruz C, Villegas-Cortez J, Sossa-Azuela JH (2013) Self-organizing natural scene image retrieval. *Expert Syst Appl* 40(7):2398–2409
2. Qin J, Yung NHC (2010) Scene categorization via contextual visual words. *Pattern Recogn* 43:1874–1888
3. Song F, Zhang D, Chen Q, W Jizhong (2007) Face recognition based a novel linear discriminant criterion. *Pattern Anal Applic* 10:165–174
4. Deng J, Zhang Z, Marchi E, Schuller B (2013) Sparse auto-encoder-based feature transfer learning for speech emotion recognition. In: *Affective computing and intelligent interaction, human association conference*, pp 511–516
5. Krizhevsky A, Hinton GE (2010) Factored 3-way restricted Boltzmann achiness for modelling natural images. In: *International conference on artificial intelligence and statistics*, pp 621–628
6. Chen Y, Xu W, Wu J, Zhang G (2012) Fuzzy maximum scatter discriminant analysis in image segmentation. *J Convergence Inf Technol* 7(5)
7. Yang J, Yang JY (2003) Why can LDA be performed in PCA transformed space. *Pattern Recognit* 2:563–566

8. Li LJ, Su H, Fei-Fei L (2007) What, where and who? Classification events by scene and object recognition. In: Eleventh IEEE international conference on computer vision, pp 1–8
9. Wu J, Rehg JM (2009) Beyond the Euclidean distance: creating effective visual codebooks using the histogram intersection kernel. In: Twelfth IEEE international conference on computer vision, pp 630–637
10. Li LJ, Su H, Lim YW, Fei-Fei L (2010) Objects as attributes for scene classification. In: European conference computer vision, pp 1–13
11. Nakayama H, Harada T, Kuniyoshi Y (2010) Global Gaussian approach for scene categorization using information geometry. In: IEEE conference on computer vision and pattern recognition, pp 2336–2343
12. Li LJ, Su H, Xing EP, Fei-Fei L (2010) Object bank: a high-level image representation for scene classification & semantic feature sparsification. In: Proceedings of the neural information processing systems
13. Zhou L, Zhou Z, Hu D (2013) Scene classification using a multi-resolution bag-of-features model. *Pattern Recogn* 46:424–433
14. Gao SH, Tsang IWH, Chia LT (2010) Kernel sparse representation for image classification and face recognition. In: European conference computer vision, pp 1–14
15. Huang GB, Lee H, Learned-Miller E (2012) Learning hierarchical representations for face verification with convolutional deep belief networks. In: IEEE Conference on computer vision and pattern recognition, pp 2518–2525
16. Lin SW, Ying KC, Chen SC, Lee ZJ (2008) Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert Syst Appl* 35 (4):1817–1824