Information Fusion 41 (2018) 16–24

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Hierarchical ELM ensembles for visual descriptor fusion

Stevica Cvetković[a,*], Miloš B. Stojanović[b], Saša V. Nikolić[a]

[a] Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, Niš 18000, Serbia
[b] College of Applied Technical Sciences Niš, Aleksandra Medvedeva 20, Niš 18000, Serbia

ARTICLE INFO

Article history:
Received 26 December 2016
Revised 9 June 2017
Accepted 27 July 2017
Available online 28 July 2017

Keywords:
Feature fusion
Extreme Learning Machine
Hierarchical classifiers
Scene classification

ABSTRACT

Extreme Learning Machines (ELM) have been successfully applied to variety of classification problems by utilizing a single descriptor type. However, a single descriptor may be insufficient for the visual classification task, due to the high level of intra-class variability coupled with low inter-class distance. Although several studies have investigated methods for combining multiple descriptors by ELM, they predominantly apply a simple concatenation of descriptors before classifying them. This type of descriptor fusion may impose problems of descriptor compatibility, high dimensionality and restricted accuracy. In this paper, we propose a hierarchical descriptors fusion strategy at the decision level ("late-fusion"), which relies on ELM ensembles (ELM-E). The proposed method, denoted as H-ELM-E, effectively combines multiple complementary descriptors by a two-level ELM-E based architecture, which ensures that a more informative descriptors will gain more impact on the final decision. In the first level, a separate ELM-E classifier is trained for every image descriptor. In the second level, the output scores from the previous level are aggregated into the mid-level representation which is conducted to an additional ELM-E classifier. The exhaustive experimental evaluation confirmed that the proposed hierarchical ELM-E based strategy is superior to the single-descriptor methods as well as "early fusion" of multiple descriptors, for the visual classification task. Additionally, it was shown that significant accuracy improvement is achieved by integrating ensembles of ELM as a basic classifier, instead of using a single ELM.

1. Introduction

In the last years, there have been great advances in natural scene image processing. The research was focusing both on the low-level tasks, such as denoising or segmentation, and high level ones, such as detection or classification. A variety of algorithms have been developed for the classification at the pixel level, however the problem becomes more complex at the level of the complete scene classification. The goal of the scene classification is to label an image according to a set of predefined semantic categories (e.g. forest, river, mountain, desert, etc.). It is a challenging problem, because of large variability within a given class in the sense of content, color, scales and orientations. High intra-class variability could be coupled with low inter-class distance, a problem that grows even more as finer classification is required. Research on natural scene classification has been focusing both on the use of suitable image descriptors and of appropriate classification algorithms. A variety of image texture descriptors have been proposed in the literature [13,37,50], and applied to scene classification. In order to make these descriptors more robust, it was found necessary to include additional visual cues, such as color information. It has been employed to improve the performance of scene classification algorithms due to the complementary characteristics among the color channels [43]. Although, there is an increasing amount of work on combining texture and color descriptors [4,7,17,27,45], effective fusion of descriptors by assessing their complementarity is still an open research problem in computer vision.

This motivated us to explore the complementary visual information in order to boost the scene classification performance. To make image descriptors more robust, we found it necessary to simultaneously include multiple visual cues (i.e. texture, color, etc.), by using appropriate fusion strategy. The fusion process can occur at the descriptor level, or at the decision level [2,44]. While descriptor-level fusion (i.e. „early fusion") integrates heterogeneous descriptors together into a single vector, decision-level fusion („late fusion") operates on output classification scores of each individual descriptor and combines them into a final decision. Despite its simplicity and computational efficiency, the early fusion approach may impose problems of descriptor compatibility, high dimensionality and restricted accuracy. The basic approach to the late fusion is to use a fixed weight for each classifier score and afterwards compute a weighted sum of the scores as the final result. This assumes that all the classifiers share the same weight and is unable to consider the differences of the classifier's individual prediction

* Corresponding author.
E-mail address: stevica.cvetkovic@elfak.ni.ac.rs (S. Cvetković).

http://dx.doi.org/10.1016/j.inffus.2017.07.003
1566-2535/

capability. Therefore, in the proposed work, we focus on the late fusion of descriptors where an additional classifier is trained to estimate the specific fusion weights for each separate descriptor. The proposed method is investigated in the context of the scene classification task.

As the basic classifier, we consider a single hidden layer feedforward neural networks (SLFN), which is an alternative to the commonly used SVM [12]. Concretely, we investigate a recently introduced SLFN training algorithm, termed as Extreme Learning Machine (ELM) [20,24]. The choice of ELM classifier is due to its extremly efficient training procedure and highly accurate classification performance. The main drawback of traditional artificial neural networks and SVM is their training speed, which has been a major issue for practical applications, especially when real-time output of the system is needed. The ELM drastically increases training speed of SLFN by randomly generating input weights and biases for hidden layer nodes, instead of iteratively adjusting their parameters by commonly used gradient-based methods. The output weights of the hidden layer are then analytically computed by a least squares method. Besides minimizing the training error, ELM finds the smallest norm of output weights and hence tends to give better generalization performance than gradient-based learning algorithms, such as backpropagation. Moreover, the ELM can "naturally" handle the multi-class classification problem with the architecture of multiple output nodes equal to the number of pattern classes. This is an advantage compared to the widely used SVM method that applies one-versus-all or one-versus-one strategy to handle non-binary cases [40]. It would be highly beneficial to study possibilities for ELM integration into the heterogeneous descriptor fusion scheme, as presented in this work. The ELM has already been applied to a variety of classification-related problems including: texture classification [26], protein sequence classification [11], remote sensing image classification [15,33], landmark recognition [8,10], etc.

Compared to existing machine learning techniques, the ELM is conceptually simpler and computationally more efficient while demonstrating high generalization capabilities. However, the random assignment of parameters introduces suboptimal input weights and biases into hidden layer that may result in unstable and non-optimal output. A natural way to overcome this drawback is to use an Ensemble of ELMs according to the established principles of randomized learners, such as Random Forest [5]. Several algorithms for the formation of ELM ensembles were recently proposed [9,34,42], including our Average Score Aggregation [14]. The main advantage of ensembles comes from the fact that combined outputs from several diverse learners can increase the generalization capabilities of a single classifier used in the ensemble [18]. To further improve diversity, several learner-independent techniques such as resampling, label switching, and feature space partitions, could be applied [3].

Inspired by the two previous trends of descriptor fusion and ELM ensembles, we propose to couple them in such a way which allows ELMs to directly select those descriptors that best discriminate the target classes, from a set of descriptor candidates. Our approach for descriptor fusion is hierarchical. We propose a two-level ELM based architecture which ensures that a more informative descriptors will gain more significance in the final decision. In the first level, a separate ELM classifier is trained for every image descriptor. Then, in the second level, the output scores returned by the first level classifiers are aggregated to obtain the mid-level representation. Mid-level descriptor is then used as an input for the second level ELM classifier, to produce the final classification result. In this way we allow a second level classifier to directly favor those descriptors that best discriminate the target classes. To further improve accuracy of the method, we propose to integrate Ensembles of ELM as a basic classifier, instead of a single ELM. Ensembles of ELM are proven to be able to improve classification accuracy largely, without significant time consumption [9,14]. In this work, we successfully integrated Ensemble ELMs in the proposed hierarchical ELM architecture for the scene classification task. As the main contribution of the paper we assume introduction of a novel descriptor fusion method that effectively tackles image intra-class diversity by proposing a hierarchical ELM based approach. Apart from the theoretical contribution, we performed extensive evaluation over the two public scene datasets which proved that the proposed algorithm can reach highly accurate results without computationally complex operations. A comparative evaluation demonstrates increased classification accuracy of the proposed H-ELM-E method compared to the accuracy when separate descriptors are used, as well as to the early fusion of descriptors (i.e. descriptor concatenation). In addition, the experiments demonstrate high level of computational efficiency of the complete scene classification pipeline.

The reminder of the paper is organized as follows. Section 2 gives a brief overview of ELM and ensembles of ELMs for multi-class classification, and then introduces the proposed method for hierarchical descriptor fusion which relies on ELM ensembles. Section 3 describes the extraction of the visual descriptors used in the proposed classification scheme. Experimental results and discussion are presented in Section 4, while Section 5 draws conclusions and proposes ideas for future work.

2. Hierarchical fusion of Extreme Learning Machines (ELM)

Fusion of classifiers aims to include mutually complementary individual classifiers which are characterized by high diversity and accuracy [47]. It is intuitive that increasing of diversity should lead to the better accuracy of the combined classifier, but there is no formal proof of this dependency. Brown et al. [6] noticed that we can successfully ensure diversity by independent generation of individual classifiers based on random techniques. The advantage of using ELMs in the fusion is that its diversity comes naturally from randomness in its hidden layer of neurons. Additional increase in diversity of the proposed hierarchical method is provided by integrating ensemble of ELMs as a basic classifier, instead of using a single ELM.

We will first give a brief overview of ELM and ensemble of ELMs, and afterwards describe the proposed hierarchical ELM-based algorithm for heterogeneous descriptor fusion (H-ELM-E).

2.1. ELM for multiclass classification

Let suppose that we have N training samples denoted as $(\boldsymbol{x}_j,\ \boldsymbol{y}_j),\ j = 1,\ldots,N$, where $\boldsymbol{x}_j = [x_{j1},\ x_{j2},\ \ldots,\ x_{jn}]^T \epsilon \boldsymbol{R}^n$ represents the j-th training sample of the dimension n, and $\boldsymbol{y}_j = [y_{j1},\ y_{j2},\ \ldots,\ y_{jm}]^T \epsilon \boldsymbol{R}^m$ represents the j-th training label of the dimension m, where m is the number of classes. In the context of the visual feature fusion, $\boldsymbol{x}_j$ could be assumed as an image descriptor, while $\boldsymbol{y}_j$ is an m dimensional binary vector of class labels, with value "1" at the position of the corresponding class, and value "0" at other positions. The output of an ELM, with L hidden neurons and activation function $h(x)$ is defined as

$$f\left(\boldsymbol{x}_j\right) = \sum_{i=1}^{L} \boldsymbol{\beta}_i h\left(\boldsymbol{w}_i \cdot \boldsymbol{x}_j + b_i\right);\ \ j = 1,\ldots,N \tag{1}$$

where $h()$ is a nonlinear piecewise continuous activation function, $\boldsymbol{\beta_i} \in \boldsymbol{R}^m$ represents the weight vector connecting the ith hidden neuron and all the output neurons, $\boldsymbol{w_i} \in \boldsymbol{R}^n$ is the weight vector connecting the ith hidden neuron and all input neurons, and b_i is the threshold of the ith hidden neuron. Although sigmoid activation function is the most commonly used in practical applications,