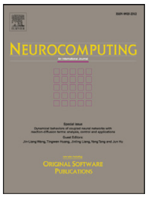
Neurocomputing 277 (2018) 4–11

Contents lists available at ScienceDirect

Neurocomputing

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



Multi-modal local receptive field extreme learning machine for object recognition

Huaping Liu*, Fengxue Li, Xinying Xu, Fuchun Sun

Department of Computer Science and Technology, Tsinghua University, State Key Lab. of Intelligent Technology and Systems, TNLIST, Beijing, PR China

ARTICLE INFO

Article history:
Received 18 April 2016
Revised 22 March 2017
Accepted 3 April 2017
Available online 23 August 2017

Keywords:
Representation learning
Multi-modal
Local receptive field
Extreme learning machine

ABSTRACT

Learning rich representations efficiently plays an important role in the multi-modal recognition task, which is crucial to achieving high generalization performance. To address this problem, in this paper, we propose an effective Multi-Modal Local Receptive Field Extreme Learning Machine (MM-LRF-ELM) structure, while maintaining ELM's advantages of training efficiency. In this structure, LRF-ELM is first conducted for feature extraction for each modality separately. And then, the shared layer is developed by combining these features from each modality. Finally, the Extreme Learning Machine (ELM) is used as supervised feature classifier for the final decision. Experimental validation on Washington RGB-D Object Dataset illustrates that the proposed multiple modality fusion method achieves better recognition performance.

1. Introduction

Object recognition is a challenging task in computer vision and important for making robots useful in home environments. With the recent advent of depth cameras, an increasing amount of visual data not only contains color but also depth measurements. Compared to RGB data, which provides information about appearance and texture, depth data contains additional information about object shape and it is invariant to lighting or color variations [1].

In recent years, various approaches that have been proposed for RGB-D object recognition: methods with hand-crafted features [2–4], and methods with learned feature [5–10]. Moreover, the classical neural network structure, like convolutional neural network networks (CNNs), is also applied to the object recognition field [24–26] and it have recently been shown to be remarkably successful for recognition on RGB images [23].

Though traditional gradient-based learning algorithms (like BP Neural network) [11] have been widely used in the training of multilayer feedforward neural networks [21,22], these gradient-based learning algorithms are still relatively slow in learning and easily get stuck in local minima [13]. Furthermore, the activation functions used in these gradient-based tuning methods need to be differentiable.

In order to overcome the drawbacks of gradient-based methods, Huang et al. proposed an efficient training algorithm for the single-hidden layer feedforward neural network (SLFN) called Extreme Learning Machine (ELM) [12,14]. It increases the learning speed by means of randomly generating input weights and hidden biases, and the output weights are determined by using Moore–Penrose (MP) generalized inverse. Compared with the traditional gradient-based learning algorithms, ELM not only learns much faster with higher generalization performance [27,30] but also avoids many difficulties faced by gradient-based learning methods such as stopping criteria, learning rate, learning epochs, and local minima. What is more, more and more deep ELM learning algorithms has been proposed [33,34] to capture relevant higher-level abstractions. However, ELM with local connections has not attracted much research attention yet. Ref. [15] has proved that the application of the local receptive fields based ELM (LRF-ELM) has better performance than conventional deep learning solutions [16,31,32] in image processing and speech recognition.

However, the aforementioned works do not refer to the multi-modal problem [28,29]. Thus, in this paper, we extend the LRF-ELM and propose a Multi-Modal LRF-ELM (MM-LRF-ELM) framework. The proposed MM-LRF-ELM is applied to multi-modal learning task, while maintaining its advantages of training efficiency. The contributions of this work are summarized as follows:

1. We propose an architecture: multi-modal LRF-ELM framework, to construct the nonlinear representation from different aspects of information sources. The important merit of such a method is that the training time is greatly shortened and the testing efficiency is highly improved.

[*] Corresponding author.
E-mail address: hpliu@tsinghua.edu.cn (H. Liu).

http://dx.doi.org/10.1016/j.neucom.2017.04.077
0925-2312/© 2017 Elsevier B.V. All rights reserved.

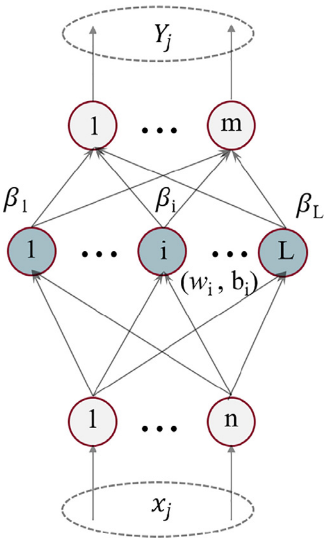


Fig. 1. The model of basic ELM.

2. We evaluate our multimodal network architecture on the Washington RGB-D Object Dataset [4]. The obtained results show that the proposed fusion method obtains rather promising results.

The remainder of this paper is organized as follows: Section 2 introduces the related works, including the fundamental concepts and theories of ELM; Section 3 describes the proposed MM-LRF-ELM framework; Section 4 compares the performance of MM-LRF-ELM with single modality and other methods; while Section 5 concludes this paper.

2. Brief review for ELM

ELM was proposed in Huang et al. [12] (Fig. 1). Suppose we are training SLFNs with K hidden neurons and activation function g(x) to learn N distinct samples $\{\mathbf{X}, \mathbf{T}\} = \{\mathbf{X}_j, \mathbf{t}_j\}_{j=1}^{N}$,where $\mathbf{x}_j \in \mathbf{R}^n$ and $\mathbf{t}_j \in \mathbf{R}^m$. In ELM, the input weights and hidden biases are randomly generated instead of tuned. By doing so, the nonlinear system has been converted to a linear system

$$\mathbf{Y}_j = \sum_{i=1}^{L} \beta_i g_i(\mathbf{x}_j) = \sum_{i=1}^{L} \beta_i g(\mathbf{w}_i^T \mathbf{x}_j + \mathbf{b}_i) = t_j,\ j = 1, 2, \ldots N \tag{1}$$

where $\mathbf{Y}_j \in \mathbf{R}^m$ is the output vector of the jth training sample, $\mathbf{W}_i \in \mathbf{R}^n$ is the input weight vector connecting the input nodes to the ith hidden node,b_i denotes the bias of the ith hidden neuron;$\beta_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{im})^T$ denotes the weight vector connecting the ith hidden neuron and output neurons; $g(\cdot)$ denotes hidden nodes nonlinear piecewise continuous activation functions. The above N equations can be written compactly as:

$$\mathbf{H}\beta = \mathbf{T} \tag{2}$$

where the matrix T is target matrix,

$$\mathbf{H} = \begin{bmatrix} g(\mathbf{w}_1^T \mathbf{x}_1 + \mathbf{b}_1) \cdots g(\mathbf{w}_L^T \mathbf{x}_1 + \mathbf{b}_L) \\ \vdots \ldots \vdots \\ g(\mathbf{w}_1^T \mathbf{x}_N + \mathbf{b}_1) \cdots g(\mathbf{w}_L^T \mathbf{x}_N + \mathbf{b}_L) \end{bmatrix} \tag{3}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{bmatrix}, \mathbf{T} = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix} \tag{4}$$

Thus, the determination of the output weights (linking the hidden layer to the output layer) is as simple as finding the least-square solution to the given linear system. The minimum norm least-square (LS) solution to the linear system (1) is

$$\hat{\beta} = \mathbf{H}^{\dagger}\mathbf{T} \tag{5}$$

where $H\dagger$ is the MP generalized inverse of matrix H. As analyzed by Huang, et al., ELM using such MP inverse method tends to obtain good generalization performance with dramatically increased learning speed.

3. Multi-modal LRF-ELM

3.1. Model architecture

Our architecture, which is depicted in Fig. 2, employs the LRF-ELM as the learning unit to learn shallow and deep information. The multi-modal training architecture is structurally divided into three separate phases: unsupervised feature representation for each modality separately, feature fusion representation and supervised feature classification.

As shown in Fig. 2, we perform feature learning to have representations of each modality (RGB and Depth) before they are mixed. Each modality is given to a single LRF-ELM net layer which provides useful translational invariance of low-level features such as edges and allows parts of an object to be deformable to some extent.

Mathematically, the output of each modality can be separately calculated. where $\mathbf{H}_1^c, \mathbf{H}_2^d \in N \times K \cdot (d - r + 1)^2$, the parameter N is the input samples, K is the number of feature maps , d is the input size and r is the size of the receptive field. $\mathbf{H}_1^c, \mathbf{H}_2^d$ are the pooling layer feature matrixes representing non-linear representations extracted from features of each modality, where c denotes the LRF-ELM I, which extracts the feature of the RGB image and d denotes the LRF-ELM II, which extracts the feature of the Depth image. In our work, each LRF-ELM net layer has the same parameters.

A single LRF-ELM net layer extracts low level features from RGB and depth images respectively. Both representations are given as input to another LRF-ELM layer, the combination process is as follows:

$$\mathbf{H} = \left[\mathbf{H}_1^c; \mathbf{H}_2^d\right]^T \tag{6}$$

Finally, the original ELM is performed to make a final decision based on the joint representation (Fig. 3). Through the proposed approach, multi-modal system can be developed as one whole system rather than being developed as separate expert systems for each modality.

3.2. Unsupervised feature representation

In this work, we adopt the local receptive fields based on ELM (LRF-ELM) to extract the features. In LRF-ELM, the links between input and hidden layer nodes are sparse and bounded by corresponding receptive fields, which are be sampled from any continuous probability distribution [15]. Fig. 4 illustrates that the process of learning representation from the features of each modality. The LRF-ELM consists of two basic operations:

(1) Generate the initial weight matrix $\hat{\mathbf{A}}_{init}^c$, $\hat{\mathbf{A}}_{init}^d$ randomly. With the input size $d \times d$ and the receptive field $r \times r$, the size of the feature map should be $(d - r + 1) \times (d - r + 1)$.

$$\begin{aligned} &\hat{\mathbf{a}}_k^c, \hat{\mathbf{a}}_k^d \in \mathbf{R}^{r^2} \\ &\hat{\mathbf{A}}_{init}^c, \hat{\mathbf{A}}_{init}^d \in \mathbf{R}^{r^2 \times k}, k = 1, 2, 3 \ldots K \end{aligned} \tag{7}$$

then, orthogonalize the initial weight matrix $\hat{\mathbf{A}}_{init}^c, \hat{\mathbf{A}}_{init}^d$,using singular value decomposition (SVD) method.