Neurocomputing 000 (2017) 1–13

Contents lists available at ScienceDirect

Neurocomputing

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

A novel multimodal retrieval model based on ELM

Yu Zhang*, Ye Yuan, Yishu Wang, Guoren Wang

School of Computer Science and Engineering, Northeastern University Shenyang 110819 China

ARTICLE INFO

Article history:
Received 29 September 2016
Revised 10 January 2017
Accepted 16 March 2017
Available online xxx

Keywords:
Extreme Learning Machine
Multimodal
Probabilistic Latent Semantic Analysis
Single hidden-layer feedforward neural networks
Modality
Regression
Multimedia

ABSTRACT

In this paper, we propose a novel multimodal retrieval model based on the Extreme Learning Machine (ELM). We exploit two multimedia modalities, the image and text, to achieve the multimodal retrieval. To begin with, we employ the probabilistic Latent Semantic Analysis (pLSA) to respectively simulate the generating processes of texts and images. So we obtain the appropriate representations of the images and those of the texts. Furthermore, ELM is used for training the correlation between the representations of the images and those of the texts. So the multimodal retrieval is implemented by the learned single-hidden layer feedforward neural networks (SLFNs). Additionally, the binary classifiers are trained to improve the accuracy of the multimodal retrieval model. This multimodal model can easily be extended into other modalities and extensive experimental results demonstrate the effectiveness and efficiency of this model based on ELM.

1. Introduction

As multimedia applications have been extended to many fields of people's lives, the demand for dealing with multimedia information presents new challenges. Traditionally, the query by the example (QBE) is widely accepted as the general multimedia retrieval mode. For example, we can search the similar images with the fixed query image in the multimedia database or on the Internet. Nowadays, multimodal retrieval is proposed and furthermore is paid more attention to in order to reply the new multimedia applications. Specifically speaking, we can find the answers of one type of the multimedia modality in terms of the query example of the other type of the multimedia modality. For instance, we can search images by query examples of texts; or we can search texts by query examples of images. These new challenges motivate researchers to present new methods and improve existing methods so as to meet the demand for multimedia applications.

To address the multimodal retrieval, in this paper, we employ the probabilistic Latent Semantic Analysis (pLSA) to propose a new multimodal retrieval model based on the Extreme Learning Machine (ELM). For convenience, in this paper we exploit two multimedia modalities, the image and text, to achieve the multimodal retrieval. That is, we search texts by the query examples of images; and we search images by the query examples of texts. However, our method can easily be extended to other modalities such as audio, video and etc.

In the traditional multimedia retrieval, the notorious semantic gap [21] badly deteriorates the performance of the retrieval. To cope with the semantic gap, the semantic analysis models are introduced into this field so as to bridge this gap [19,20,22]. Thus the semantic analysis models have been extensively used in the multimedia applications [28,29,31,33,36]. The probabilistic Latent Semantic Analysis (pLSA) is one of the semantic analysis models and it employs the probabilistic semantic method to transform the textual document into the multi-dimensional vector representation [19].

The feedforward neural networks have been widely used in many fields of feature learning, classification, regression, compression and etc [12]. But it has two major bottlenecks: the learning algorithms is too slow and its parameters must be tuned iteratively [2]. The Extreme Learning Machine (ELM) is proposed to address the problems of single-hidden layer feedforward neural networks (SLFNs) [1,8]. ELM is a learning algorithm that can fast learn the parameters of the SLFNs [3,4]. Nowadays, ELM has been extensively applied to more and more fields [6,7,10,14–17].

In the proposed multimodal retrieval model based on ELM, firstly, we exploit the semantic analysis method – pLSA to stimulate the generative processes of the textual documents in the training set and we get the latent aspects(topics) of these textual documents by EM methods. Secondly, we assign the feature vectors of the training images into different clusters and we consider each cluster as one visual word. Thus each image can be regarded as

* Corresponding author.
E-mail address: 2145640646@qq.com (Y. Zhang).

http://dx.doi.org/10.1016/j.neucom.2017.03.095
0925-2312/

one document that consists of some visual words. Then we also employ pLSA to simulate the generative processes of the training images. Each image has been matched with one text in the training set, so intuitively the image has the same semantic meanings with its matched text [23,24,26,27,30,37]. Thirdly, we employ the SLFNs to analyze the correlation between the latent aspects of images and texts based on their same semantic senses. Moreover, ELM is applied to learning the regression of the correlation between the semantic representations of images and texts. When one query image arrives, we replace its feature vectors with the corresponding visual words and use pLSA to obtain its semantic representation of the latent aspects. Furthermore, we transform its latent aspects of image into its textual latent aspects by the SLFNs. Therefore, we can search the texts by the example of the query image. At the same time, if one query text arrives, we can implement the image retrieval in the similar process with the image query. In the addition, the binary word classifiers based on ELM are learned to verify whether the candidate words are related to the query image, which can improve the accuracy of the multimodal retrieval model. Generally, this multimodal retrieval model based on ELM can be extended to more modalities such as audio, video and etc. by using the suitable semantic representations of different modalities.

This paper is organized as follow. Section 2 introduces the preliminary. In Section 3, we present the multimodal retrieval model based on ELM. Section 4 details the training and the inference. The binary word classifier is introduced in Section 5. The experimental results are shown in Section 6. In Section 7, we provide the related works. Finally, we conclude the paper and present the future works in Section 8.

2. Preliminary

In this section, we make the overviews of Extreme Learning Machine (ELM) and the standard probabilistic Latent Semantic Analysis (pLSA).

2.1. Extreme Learning Machine

Before Extreme Learning Machine (ELM) was proposed, there are no effective learning algorithms to fast learn the parameters of feedforward neural networks. The traditional methods are very slow because the parameters of the neural networks are tuned iteratively [13,18]. However, ELM is originally proposed to dramatically reduce the learning time of the single-hidden layer feedforward neural networks (SLFNs) [1,3]. It is a learning algorithm for SLFNs and it has not only faster learning speed but also better generalization performance than the traditional learning algorithms [5,9,11].

Given N arbitrary samples $(x_i,\ t_i)$, where $i = 1, 2, \ldots, N$, $x_i = [x_{i1}, x_{i2}, \ldots, x_{in}]^{\mathrm{T}} \in R^n$ and $t_i = [t_{i1}, t_{i2}, \ldots, t_{im}]^{\mathrm{T}} \in R^m$, the output function of generalized SLFNs with L hidden function and with activation function $G(x)$ is

$$f_L(x_j) = \sum_{i=1}^{L} \beta_i G_i(x_j) = \sum_{i=1}^{L} \beta_i G(a_i, b_i, x_j) = o_j,\ j = 1, 2, \ldots, N \quad (1)$$

where the hidden node parameters $a_i = [a_{i1}, a_{i2}, \ldots, a_{in}]^{\mathrm{T}}$ and b_i respectively are the weight vector and the threshold, and specifically, a_i connects ith hidden node and the input nodes and b_i is the threshold of the ith hidden node; the weight vector $\beta_i = [\beta_{i1}, \beta_{i2}, \ldots, \beta_{im}]^{\mathrm{T}}$ connects ith hidden node and the output nodes [2].

ELM has proved that there exist SLFNs with zero error for N samples. That is, $\sum_{j=1}^{N} \|o_j - t_j\| = 0$ [10]. Furthermore, ELM can learn the parameters of the SLFNs in the different way from the traditional methods [13,18].

To begin with, ELM proves that the hidden node parameter sequence $\{a_i, b_i\}_{i=1}^{L}$ can be randomly generated. Then the hidden layer output function is:

$$h(x) = [G(a_1, b_1, x), G(a_2, b_2, x), \ldots, G(a_L, b_L, x)] \quad (2)$$

Moreover, the hidden layer output matrix can be written as:

$$\mathbf{H} = [h(x_1), h(x_2), \ldots, h(x_N)]^{\mathrm{T}} \quad (3)$$

So, Eq. (4) can be obtained by putting Eqs. (1)–(3) into $\sum_{j=1}^{N} \|o_j - t_j\| = 0$.

$$\mathbf{H}\beta = \mathbf{T} \quad (4)$$

where

$$\mathbf{H} = \begin{pmatrix} G(a_1, b_1, x_1) \ldots G(a_L, b_L, x_1) \\ \ldots \\ G(a_1, b_1, x_N) \ldots G(a_L, b_L, x_N) \end{pmatrix}_{N \times L} \text{ and}$$

$$\beta = \begin{pmatrix} \beta_1^{\mathrm{T}} \\ \vdots \\ \beta_L^{\mathrm{T}} \end{pmatrix}_{L \times m} = \begin{pmatrix} \beta_{11} \ldots \beta_{1m} \\ \ldots \\ \beta_{L1} \ldots \beta_{Lm} \end{pmatrix}_{L \times m} \text{ and}$$

$$\mathbf{T} = \begin{pmatrix} t_1^{\mathrm{T}} \\ \vdots \\ t_N^{\mathrm{T}} \end{pmatrix}_{N \times m} = \begin{pmatrix} t_{11} \ldots t_{1m} \\ \ldots \\ t_{N1} \ldots t_{Nm} \end{pmatrix}_{N \times m}$$

If the number of the hidden nodes L is equal to the number of training samples N, for the randomly generated parameter sequence of the hidden nodes $\{a_i, b_i\}_{i=1}^{L}$, β can be exactly learned. That is, SLFNs can be solved with zero error [2].

However, in most cases L is far less than N, namely, $L \ll N$. So β may not be exactly obtained so that $\mathbf{H}\beta = \mathbf{T}$, as for the randomly generated parameter sequence $\{a_i, b_i\}_{i=1}^{L}$. But ELM can employ the smallest norm least squares solution to approximate β so that $\|\mathbf{H}(a, b, x)\beta - \mathbf{T}\|$ has the smallest error. Then

$$\beta = \mathbf{H}^{\dagger}\mathbf{T} \quad (5)$$

where $\mathbf{H}\dagger$ is the Moore–Penrose generalized inverse of matrix $\mathbf{H}$.

As for the binary classification case, ELM presents the output through the single output node. On the other hand, with respect to the problem of the multi-class classifier, ELM also can approximate any target continuous functions and the output of the ELM classifier is approximated to the class labels.

Given m-class, classifiers have m output nodes. The multi-class classification problem can be formulated as:

$$Minimize: L_{p_{ELM}} = \frac{1}{2}||\beta||^2 + C\frac{1}{2}\sum_{i=1}^{N}||\xi_i||^2$$

$$subject to: h(x_i)\beta = t_i^{\mathrm{T}} + \xi_i^{\mathrm{T}}, \forall i = 1, \ldots, N \quad (6)$$

where C is user-specified parameter and $\xi_i = [\xi_{i1}, \ldots, \xi_{im}]^{\mathrm{T}}$ is the training error vector of the training sample t_i.

Moreover, ELM approximate the output of the multi-class classification by addressing the dual optimization problem:

$$L_{D_{ELM}} = \frac{1}{2}\|\beta\|^2 + C\frac{1}{2}\sum_{i=1}^{N}\|\xi_i\|^2 - \sum_{i=1}^{N}\sum_{j=1}^{m}\alpha_{ij}(h(x_i)\beta_j - t_{ij} - \xi_{ij}) \quad (7)$$

where α_{ij} is the Lagrange multiplier and it corresponds to the training sample t_{ij}; β_j is the weight vector connecting the hidden layer and the jth output node and $\beta = [\beta_1, \ldots, \beta_m]$.

Please cite this article as: Y. Zhang et al., A novel multimodal retrieval model based on ELM, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.03.095