Improving The Multimodal Probabilistic Semantic Model by ELM Classifiers

Yu Zhang, Ye Yuan, Fangda Guo, Yishu Wang, and Guoren Wang

Northeastern University
Shenyang, 110819, China

Abstract. The multi-modal retrieval is considered as performing information retrieval among different modalities of multimedia information. Nowadays, it becomes increasingly important in the information science field. However, it is so difficult to bridge the meanings of different multimedia modalities that the performance of multimodal retrieval is deteriorated now. In this paper, we propose a new mechanism to build the relationship between visual and textual modalities and to verify the multimodal retrieval. Specifically, this mechanism depends on the multimodal binary classifiers based on the Extreme Learning Machine(ELM) to verify whether the answers are related to the query examples. Firstly, we propose the multimodal probabilistic model to rank the answers according to their generative probabilities. Furthermore, we build the multimodal binary classifiers to filter out unrelated answers. The multimodal binary classifiers are called the word classifiers. It can improve the performance of the multimodal probabilistic semantic model. The experimental results show that the multimodal probabilistic semantic model and the word classifiers are effective and efficient. Also they demonstrate that the word classifiers based on ELM not only can improve the performance of the probabilistic semantic model but also can be easily applied to other probabilistic semantic models.

Keywords: Extreme Learning Machine, Multimodal, Classifier, Probabilistic semantic model, probabilistic Latent Semantic Analysis, Single hidden-layer feedforward neural networks, Retrieval

1 Introduction

Nowadays, with the rapid development of the information technology not only the scale but also the type of multimedia information has explosively increased. So the multimedia information usually has the multimodal nature. Specifically, the multimedia document consists of a variety of different modalities of multimedia information. Meanwhile the different modalities of multimedia information in the same multimedia document generally contain the same semantic senses [25] [29] [30] [32] [34] [35]. These changes bring new challenges to the information retrieval and the multimedia database. How to effectively and efficiently search multimodal information in the multimedia database is the new focus of the information retrieval, multimedia and database fields. The multimodal retrieval need exploit the integrated analysis of different modalities of multimedia information so as to obtain the potential correlation of different modalities. Then it can employ the potential correlation to achieve different kinds of the multimodal retrieval. While traditional multimedia retrieval methods depend on the query by the example(QBE), the multimodal retrieval can search other kinds of modalities of multimedia information by different modalities of query examples. For examples, we can search images by examples of texts; and we can search texts by examples of images.

The semantic models are extended to multimodal retrieval fields in order to bridge the correlation of different modalities of multimedia information [20] [25] [34] [35]. At the beginning the semantic models are proposed to cope with the notorious semantic gap [1] [2] [4] [19]. Then they are extensively applied to many different fields [22] [23] [36]. Moreover, the probabilistic Latent Semantic Analysis (pLSA) introduces the probabilistic model into the semantic analysis and employs the probabilistic distribution of the latent aspects to represent the document [1].

However, the probabilistic semantic methods only rank the answers according to their generative probabilities without considering other factors. In fact the answer to the query is just a *Yes/No* decision for the candidates. Hence in this paper we employ the Extreme Learning Machine (ELM) to build the binary classifiers to verify the candidates ranked by the probabilistic semantic methods in order to improve the effectiveness of the multimodal retrieval. For convenience sake, in this paper we exploit two multimedia modalities, the image and text, to achieve the multimodal retrieval.

The Extreme Learning Machine (ELM) is an effective and efficient learning algorithms for the single-hidden layer feedforward neural networks (SLFNs) [5] [9]. The feedforward neural networks have been widely and effectively applied to many fields such as feature learning, classification, regression, compression and etc. [12]. However, the traditional learning algorithms train the feedforward neural networks in the light of iteratively tuning the parameters [13] [39]. Hence the traditional algorithms are less efficient to learn the feedforward neural networks. Recently, ELM is presented to tackle these problems. It can not only fast learn SLFNs but also guarantee high training and testing accuracy. The Extreme Learning Machine further boosters the development of the feedforward neural networks because it solves the fundamental problem of the SLFNs [8] [11] [41]. Multiple kernel learning is used in ELM to optimize the choice of kernels so as to improve the performance of ELM [43] [44] [45] [46]. Nowadays, because of the excellent efficiency the ELM has been extended into a variety of different fields [7] [14] [15] [17] [42] [47] [48] [49].

To address the new challenges of the multimodal retrieval and to complement the probabilistic semantic models, in this paper, we propose a new multimodal retrieval model. This multimodal retrieval model consists two part: the multimodal probabilistic semantic model and the word classifiers based on ELM. The image feature data are continuous and the standard pLSA can be only applied to discrete data. The standard pLSA model can not straightforwardly simulate the generative process of images. It is generally assumed that the image feature data follow the multivariate Gaussian distribution under the given parameters [3] [25]. So the multivariate Gaussian distribution can be introduced to handle the continuous data such as image feature vectors. Firstly, we employ multivariate Gaussian pLSA to simulate the generative process of the images in the training set and obtain the probability distributions of the latent aspects (topics) of the images. Moreover, we assume that the images and the texts share the common latent aspects and consider the probability distributions of images as the probability distributions of the texts. Additionally, we employ the standard pLSA to simulate the generative process of the training texts under the condition that the probability distributions of the textual latent aspects are fixed. So we can obtain the probability distributions of the vocabulary under the condition of the textual latent aspects. At the same time, we consider the training image feature vectors as the input and employ ELM to train the multimodal binary classifier of every textual word in the vocabulary. These multimodal binary classifiers are called the word classifiers. For each textual word in the vocabulary, its word classifier verifies whether the image feature vector belongs to its class or not. If any feature vector of the image output true for one word, the image is related to that word. Or else, it does not. When one query image arrives, we can search its related texts in the light of the probabilistic semantic model and filter out the unrelated texts through the word classifiers. On the other hand, we employ the standard pLSA to simulate the generative process of the training texts and consider their textual latent aspects as the aspects of images. Thus we use multivariate Gaussian pLSA to simulate the generative process of the images. When one text arrives, we can search its related images by the probabilistic semantic model and filter out the unrelated images by the word classifiers. More importantly, the word classifiers based on ELM can be extensively applied to other multimodal probabilistic semantic models. So it is not limited to the proposed multimodal probabilistic semantic model in this paper. The extensive experimental results show the effectiveness and efficiency of the word classifiers based on ELM and the multimodal retrieval model. Also the experimental results demonstrate the expandability of the word classifiers. Generally speaking, this multimodal retrieval model based on ELM classifiers can be easily extended to other modalities such as audio, video and etc.

This paper is organized as follow. The section 2 introduces the overview of the ELM method. In the section 3, we propose the multimodal probabilistic semantic model. In the section 4, based on ELM we present the word classifiers that verify whether the candidate words are related to the query images. The experimental results are shown in the section 5. The section 6 states the related works. We conclude the paper and provide the future works in the section 7.

2 The Overview of Extreme Learning Machine

The traditional learning algorithms usually train the feedforward neural networks with less efficiency. So the Extreme Learning Machine (ELM) is proposed to solve the problem. The ELM is an excellently efficient and effective learning algorithm because it can fast learn the parameters of the single-hidden layer feedforward neural networks (SLFNs) with satisfactory accuracy. ELM randomly provides the values of the parameters of the hidden nodes and then learn the weight parameters that connect the hidden nodes and the output nodes. Specifically, given N arbitrary distinct samples (x_i, t_i), where $i = 1, 2, ..., N$, $x_i = [x_{i1}, x_{i2}, ..., x_{in}]^{\mathrm{T}} \in R^n$ and $t_i = [t_{i1}, t_{i2}, ..., t_{im}]^{\mathrm{T}} \in R^m$, the output function of SLFNs with L hidden function and with activation function $G(x)$ is

$$f_L(x_j) = \sum_{i=1}^{L} \beta_i G(a_i, b_i, x_j) = \sum_{i=1}^{L} \beta_i g(a_i \cdot x_j + b_i) = o_j, j = 1, 2, ..., N \tag{1}$$

where $a_i = [a_{i1}, a_{i2}, ..., a_{in}]^{\mathrm{T}}$ and b_i are the hidden node parameters, and a_i is the weight vector that connects ith hidden node and the input nodes, and b_i is the threshold of the ith hidden node; $\beta_i = [\beta_{i1}, \beta_{i2}, ...\beta_{im}]^{\mathrm{T}}$ is the weight vector and it connects ith hidden node and the output nodes [6].

For the additive nodes, in the Equation (1) the function g is the output function of the hidden nodes and $a_i \cdot x_j$ denotes the inner product of a_i and x_j. On the other hand, with respect to the Radial Basis Function (RBF) nodes, the output function of the hidden nodes can be written as the Equation (2).

$$f_L(x_j) = \sum_{i=1}^{L} \beta_i G(a_i, b_i, x_j) = \sum_{i=1}^{L} \beta_i g(b_i \left\| x_j \; - \; a_i \right\|), j = 1, 2, ..., N \tag{2}$$

ELM has proved that the hidden node parameter sequence $\{a_i, b_i\}_{i=1}^{L}$ can be randomly generated and β need be learned. The hidden layer output matrix can be written as:

$$\mathbf{H} = \begin{pmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{pmatrix} = \begin{pmatrix} G(a_1, b_1, x_1) G(a_2, b_2, x_1) ... G(a_L, b_L, x_1) \\ ... \\ G(a_1, b_1, x_N) G(a_2, b_2, x_N) ... G(a_L, b_L, x_N) \end{pmatrix}_{N \times L} \tag{3}$$

In fact, $h(x_i)$ is the random feature mapping which maps the input sample x_i from n-dimensional input space to the L-dimensional ELM feature space. So the output function of SLFNs in the Equation (1) can be written as the Equation (4).

$$\mathbf{H}\beta = \mathbf{T} \tag{4}$$

where

$$\beta = \begin{pmatrix} \beta_1^{\mathrm{T}} \\ \vdots \\ \beta_L^{\mathrm{T}} \end{pmatrix}_{L \times m} = \begin{pmatrix} \beta_{11} \beta_{12} ... \beta_{1m} \\ ... \\ \beta_{L1} \beta_{L2} ... \beta_{Lm} \end{pmatrix}_{L \times m}$$

and

$$\mathbf{T} = \begin{pmatrix} t_1^{\mathrm{T}} \\ \vdots \\ t_N^{\mathrm{T}} \end{pmatrix}_{N\times m} = \begin{pmatrix} t_{11}t_{12}...t_{1m} \\ ... \\ t_{N1}t_{N2}...t_{Nm} \end{pmatrix}_{N\times m}$$

ELM has proved that SLFNs can be learned with zero error when the number of the hidden nodes L is equal to the number of training samples N. That is, $\sum_{j=1}^{N} \|o_j - t_j\| = 0$ [6] [7]. However, usually L is far less than N, namely $L \ll N$. In this case SLFNs can not be solved with zero error. Specifically, β can not be exactly learned so that $\mathbf{H}\beta = \mathbf{T}$. But ELM exploits the smallest norm least squares solution to approximate β so that $\|\mathbf{H}\beta = \mathbf{T}\|$ has the smallest error. So when $L \ll N$, the output weights β can be calculated as Equation (5)

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

where $\mathbf{H}^{\dagger}$ is the MooreCPenrose generalized inverse of matrix $\mathbf{H}$.

In the binary classification case, ELM only employs single output node to present the predicted class label and this paper mainly concerns the binary classification problem. However, with respect to the multi-class classifier, ELM can approximate any target continuous functions. So $h(x_j)\beta$, the output of the ELM classifier, need be approximated to the class labels. Given m-class of classifiers have m output nodes. The multi-calss classification problem can be formulated as:

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

$$subjectto : h(x_i)\beta = t_i^{\mathrm{T}} + \xi_i^{\mathrm{T}}, \forall i = 1, ..., N \tag{6}$$

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

So ELM can solve multi-class classification problem by tackling the dual optimization problem:

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

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 $\beta = [\beta_1, ..., \beta_m]$.

3 The Multimodal Probabilistic Semantic Model

In this section, we describe the multimodal probabilistic semantic model in details. We assume that each multimodal document consists of one image and its associated text and that they have the same sematic senses [25] [29] [30] [32] [34] [35]. At the same time, we assume that the image and text in the multimodal document are respectively generated independently. But they share the common latent semantic aspects (topics). Therefore, if we have an image (text) in the multmodal document, we can generate its corresponding text (image) in the same multimodal document based on their same semantic senses. Namely, we can transform the semantic senses from one modality to the other modality. This section is organized as follow. Firstly, we introduce the standard pLSA. Then we train the transformation from the image modality to the textual modality so as to implement the image query. Finally, we train the transformation from the textual modality to the image modality in order to perform the text-based image retrieval.

3.1 probabilistic Latent Semantic Analysis

The probabilistic Latent Semantic analysis (pLSA) model is proposed in 2001 [1]. It employs the probabilistic semantic method to reduce the dimensions of tradition representations of text document based on the Latent Semantic Indexing (LSI). It is a generative model with the assumptions that the words are generated independently of the documents and it also assumes that the document is the collection of unorder words, that is, bag of words model (bog). It considers the document as the mixture of the latent aspects (topics) so the document can be represented as the probability distribution of the latent aspects. The standard pLSA simulates the generative process of the training documents and employs the Expectation Maximization (EM) method to maximize the probability of the creation of the training documents. In this process, the standard pLSA can learn the probability distributions of the latent aspects of the training documents and it considers them as the representations of the training documents. Also it can learn the probability distributions of the vocabulary conditioned on the latent aspects so as to calculate the representations of the testing documents. So when the testing document comes, its representation of the probability distribution of the latent aspects can be calculate based on the learned probability distributions of the vocabulary. The number of latent aspects is far less than that of words in vocabulary, so the standard pLSA representations of the document have much less dimension than tradition representations of the document. So the pLSA can be used for reducing the dimensions of the traditional representations of the documents without loss of the semantic meanings. Especially, it can bridge the notorious semantic gap [19]. Meanwhile, pLSA can be used for the document clustering and similarity of documents [23]. Recently, pLSA is extended to the multimedia field and the computer vision field [36] [37].

3.2 The Generative Process of Texts From The Visual Query

We implement the image query and then obtain the textual answers, so in the training stage we assume that the training images are generated at first and then the training texts are created based on their shared latent aspects. We employ pLSA to simulate the generative process of the training image. The standard pLSA only use discrete data and it can not directly use continuous image feature vectors. Therefore pLSA can not be straightforwardly applied to the generative process of the training image. Generally, it is assumed that the image feature vectors obey the multivariate Gaussian distribution [3] [25]. So the multivariate Gaussian distribution is introduced into pLSA to simulate the generative process of the training image. Firstly, an image is picked. Then one latent aspect is sampled conditioned on this image. So we can obtain the mean and the covariance matrix of the multivariate Gaussian distribution corresponding to that latent aspect. Furthermore, one image feature vector is sampled conditioned on the mean and the covariance matrix of this latent aspect. Repeat the generative process of the latent aspects and the generative process of the image feature vector until the training image is completed. Thus we can obtain the proportion of the latent aspects of this image. This image and its associated texts share the common semantic senses, so the proportion of the latent aspects of the associated text is fixed as that of this image when pLSA is applied to the generative process of the associated text. That is, the associated text is pick and its probability proportion of the latent aspects is fixed. Furthermore, one latent aspect is sampled based on the fixed proportion of the latent aspects. In the addition, one textual word is sampled conditioned this latent aspect. Finally, repeat the generative process of the latent aspects based on the fixed proportion and the generative process of the textual words until the associated text is completed.

Formally, the generative process of the text from the image is given as follows:

1 For the generative process of the image:

(1)Pick an image from the multimodal document d_i with prior probability $P(d_i), i \in \{1, ..., D\}$;

(2)Sample one of L common latent aspects z_l with probability $P(z_l|d_i), l \in \{1, ..., L\}$;

(3)Generate one image feature vector f_n with probability $P(f_n|z_l)$ from a multivariate Gaussian distribution $\aleph(\mu_l, \Sigma_l)$ conditioned on the z_l factor;

2 For the generative process of the text:

(1)The probability proportion $P(z|d_i)$ of the latent aspects of the text is fixed as that of the image in the same multmodal document $P(d_i)$;

(2)Sample one of L latent aspects z_l with probability $P(z_l|d_i), l \in \{1, ..., L\}$ based on the probability distribution $P(z|d_i)$;

(3)Sample one of M textual words w_m with probability $P(w_m|z_l)$ from a multinomial distribution Mult(Θ) conditioned on the z_l factor, $m \in \{1, ..., M\}$;

Generally, the image consists of a few visual feature vectors that is easily applied to clustering, comparison, classification and etc. in the computer vision and multimedia fields. It is usually assumed that these feature vectors are generated independently and that they obey the multivariate Gaussian distribution. Specifically, there are L Gaussian distributions corresponding to L visual latent aspects, z_l, $l \in \{1, ..., L\}$. And the image feature vectors are sampled from L Gaussian distributions, namely L visual latent aspects. For one fixed latent aspect z_l, the probability density function of the image feature vector f_n is written to:

$$P(f_n|z_l) = \frac{1}{(2\pi)^{C/2}|\varSigma_l|^{1/2}} e^{-\frac{1}{2}(f_n-\mu_l)^{\mathrm{T}}\varSigma_l^{-1}(f_n-\mu_l)} \tag{8}$$

where C is the dimensionality of the feature f_n, $\sum_l$ and μ_l respectively are the covariance matrix and mean of the latent aspect z_l. The probability $P(f_n|z_l)$, $l \in (1, ..., L)$ can be calculated when f_n, $\sum_l$ and μ_l are given.

For the image modality, pLSA with multivariate Gaussian is applied to the generation of the image and the log-likelihood function L_v is given by:

$$L_v = \sum_{i=1}^{D}\sum_{n=1}^{N} n(d_i, f_n)[\log P(d_i) + \log \sum_{l=1}^{L} P(f_n|z_l)P(z_l|d_i)] \tag{9}$$

where d_i is the ith multimodal document, and L is the number of the shared latent aspects of the images and their associated texts, and $n(d_i, f_n)$ is the number of the image feature vector f_n in the multimodal document d_i. Taking into consideration picking the feature vectors from the continuous multivariate Gaussian distribution, we hardly sample the same feature vectors in real practice. Thus if f_n belongs to the multimodal document d_i, the number of the feature vector f_n is always 1, namely $n(d_i, f_n) = 1$; otherwise, the number of the feature vector f_n is always 0, namely $n(d_i, f_n) = 0$. Meanwhile, the number of different features in the visual vocabulary is just the number of all features in all documents. Therefore, the Equation (9) can be written as:

$$L_v = \sum_{i=1}^{D}\sum_{n=1}^{N_i}[\log P(d_i) + \log \sum_{l=1}^{L} P(f_n|z_l)P(z_l|d_i)] \tag{10}$$

where N_i denotes the number of the image feature vectors in the multimodal document d_i.

EM algorithm is introduced to maximize the log-likelihood function L_v so as to determine the unobservable parameters $\sum_l$, μ_l and $P(z_l|d_i)$.

The E-step:

By applying Bayesian formula, one can obtain:

$$P(z_l|d_i, f_n) = \frac{P(f_n|z_l)P(z_l|d_i)}{\sum\limits_{l=1}^{L} P(f_n|z_l)P(z_l|d_i)} \tag{11}$$

The M-step:

The expectation of $E(L_v)$ is given by:

$$E(L_v) = \sum_{i=1}^{D}\sum_{n=1}^{N_i}\sum_{l=1}^{L} P(z_l|d_i, f_n)\log[P(f_n|z_l)P(z_l|d_i)] \tag{12}$$

$P(d_i)$ in Equation (10) can be calculated independently, so it is eliminated.

By maximizing Equation (12) and by combining the Equation (11), one can obtain:

$$\mu_l = \frac{\sum\limits_{i=1}^{D}\sum\limits_{n=1}^{N_i} P(z_l|d_i, f_n) f_n}{\sum\limits_{i=1}^{D}\sum\limits_{j=1}^{N_i} P(z_l|d_i, f_j)} \tag{13}$$

$$\Sigma_l = \frac{\sum\limits_{i=1}^{D}\sum\limits_{n=1}^{N_i} P(z_l|d_i, f_n)(f_n - \mu_l)(f_n - \mu_l)^{\mathrm{T}}}{\sum\limits_{i=1}^{D}\sum\limits_{j=1}^{N_i} P(z_l|d_i, f_j)} \tag{14}$$

$$P(z_l|d_i) = \frac{\sum\limits_{n=1}^{N_i} P(z_l|d_i, f_n)}{N_i} \tag{15}$$

where N_i denotes the number of image feature vectors in the document d_i. The Equations (11) in the E step and the Equation (13), (14), (15) in the M step are alternatively iterated until the convergence condition is satisfied. Thus the parameters $\sum_l$, μ_l and $P(z_l|d_i)$ can be determined.

Furthermore, the standard pLSA is employed to simulate the generative process of the text in the same multmodal document d_i with the image. But the proportion of the latent aspects of the text is fixed as that of the image, namely $P(z|d_i)$ that is the collection of $P(z_l|d_i)$, $l = 1, 2, ..., L$.

We also have assumed that the textual words are generated independently. At the same time, the log-likelihood function L_t of the text is written as [1]:

$$L_t = \sum_{i=1}^{D}\sum_{m=1}^{M} n(d_i, w_m)[\log P(d_i) + \log \sum_{l=1}^{L} P(w_m|z_l)P(z_l|d_i)] \tag{16}$$

where d_i is the ith multimodal document, and l is the number of the common latent aspects of the images and their associated texts, and $n(d_i, w_m)$ is the number of the words w_m in the text in the multimodal document d_i.

When $P(z_l|d_i)$ is fixed, $P(w_m|z_l)$ can be determined by using EM algorithm to maximize the log-likelihood function L_t [1].

The E-step:

Applying Bayesian formula, then:

$$P(z_l|d_i, w_m) = \frac{P(w_m|z_l)P(z_l|d_i)}{\sum\limits_{l=1}^{L} P(w_m|z_l)P(z_l|d_i)} \tag{17}$$

The M-step:

The expectation of $E(L_t)$ is:

$$E(L_t) = \sum_{i=1}^{D} \sum_{m=1}^{M} n(d_i, w_m) \sum_{l=1}^{L} P(z_l|d_i, w_m) \log[P(w_m|z_l)P(z_l|d_i)] \tag{18}$$

$P(d_i)$ in the Equation (16) can be calculated independently, and therefore it is eliminated. By maximizing the expectation of L_t, when $P(z_l|d_i)$ is fixed $P(w_m|z_l)$ is given by:

$$P(w_m|z_l) = \frac{\sum\limits_{i=1}^{D} n(d_i, w_m)P(z_l|d_i, w_m)}{\sum\limits_{j=1}^{M} \sum\limits_{i=1}^{D} n(d_i, w_j)P(z_l|d_i, w_j)} \tag{19}$$

The parameter $P(z_l|d_i, w_m)$ in the Equation (17) in the E step and $P(w_m|z_l)$ in the Equation (9) in the M step are alternatively iterated until the convergence condition is satisfied. Therefore, we obtain the probability distribution of the vocabulary conditioned on the common latent aspects.

When a query image arrives, we can transform it into the proportion of the common latent aspects and employ the probability distribution of the vocabulary conditioned on the common latent aspects to obtain the generative probability of each textual word in its matched text.

3.3 The Generative Process of Images From The Textual Query

Similarly, with regard to the generative process of images from the textual query, in the training stage we assume that firstly the training texts are generated and what is more the training images are generated based on their common latent aspects. We also employ the standard pLSA to simulate the generative process of the training text. Specifically, firstly, a text is picked. Secondly, one latent aspect is sampled conditioned on the text. Furthermore, one textual word is picked conditioned on this latent aspect. Repeat the generative process of the latent aspects and the generative process of the textual word until the text is completed. So the proportion of the latent aspects of the text is fixed. The image and its associated text in the same multimodal document share the common semantic senses, so they have the same proportion of the common latent aspects. The multivariate Gaussian is introduced into pLSA model because the image feature vectors are continuous. Thus we employ the multivariate Gaussian pLSA to simulate the generative process of the image under the condition that the proportion of the common latent aspects of the image is fixed. Specifically, the corresponding image is picked. Furthermore, one latent aspect is sampled based on the fixed proportion of the latent aspects. Additionally, one image feature vector is generated conditioned on the mean and the covariance matrix of this sampled latent aspect. Finally, repeat the generative process of the latent aspects based on the fixed proportion and the generative process of the image feature vector until the image is completed.

Formally, the generative process of the image from the text is given as follows:

1 For the generative process of the text:

(1)Pick a text from the multimodal document d_i with prior probability $P(d_i), i \in \{1, ..., D\}$;

(2)Sample one of K common latent aspects z_k with probability $P(z_k|d_i), k \in \{1, ..., K\}$;

(3)Sample one textual word w_m with probability $P(w_m|z_k)$ from a multinomial distribution Mult(Φ) conditioned on the z_k factor, $m \in \{1, ..., M\}$

2 For the generative process of the image:

(1)The probability proportion $P(z|d_i)$ of the latent aspects of the image is fixed as that of the text in the same multmodal document $P(d_i)$;

(2)Sample one of K latent aspects z_k with probability $P(z_k|d_i), k \in \{1, ..., K\}$ based on the probability distribution $P(z|d_i)$;

(3)Generate one image feature vector f_s with probability $P(f_s|z_k)$ from a multivariate Gaussian distribution $\Psi(\mu_k, \Sigma_k)$ conditioned on the z_k factor;

We have assumed that the textual words are generated independently, and therefore the log-likelihood function L_t of the text is given by the Equation (20).

$$L_t = \sum_{i=1}^{D} \sum_{m=1}^{M} n(d_i, w_m)[\log P(d_i) + \log \sum_{k=1}^{K} P(w_m|z_k)P(z_k|d_i)] \tag{20}$$

The probability distribution of the vocabulary conditioned on the latent aspects and that of latent aspects can be determined by using EM algorithm to get the maximum the log-likelihood function L_t [1].

The E-step:

After applying Bayesian formula, we can obtain the Equation (21).

$$P(z_k|d_i, w_m) = \frac{P(w_m|z_k)P(z_k|d_i)}{\sum\limits_{k=1}^{K} P(w_m|z_k)P(z_k|d_i)} \tag{21}$$

The M-step: The expectation of $E(L_t)$ is the Equation (22).

$$E(L_t) = \sum_{i=1}^{D} \sum_{m=1}^{M} n(d_i, w_m) \sum_{k=1}^{K} P(z_k|d_i, w_m) \log[P(w_m|z_k)P(z_k|d_i)] \tag{22}$$

By maximizing the expectation of L_t, $P(z_k|d_i)$ and $P(w_m|z_k)$ are determined by the Equation (23) and (24)

$$P(z_k|d_i) = \frac{\sum\limits_{j=1}^{M} n(d_i, w_j)P(z_k|d_i, w_j)}{n(d_i)} \tag{23}$$

$$P(w_m|z_k) = \frac{\sum\limits_{i=1}^{D} n(d_i, w_m) P(z_k|d_i, w_m)}{\sum\limits_{j=1}^{M} \sum\limits_{i=1}^{D} n(d_i, w_j) P(z_k|d_i, w_j)} \tag{24}$$

where $n(d_i)$ is the number of the textual words in the text in the multimodal document d_i. The parameter $P(z_k|d_i, w_m)$ in the Equation (21) in the E step, and $P(z_k|d_i)$ in the Equation (23) and $P(w_m|z_k)$ in the Equation (24) in the M step are alternatively iterated until the convergence condition is satisfied. Thus we obtain the proportion of the latent aspects of the texts, $P(z|d_i)$ that is the collection of $P(z_k|d_i)$. Meanwhile, the probability distribution of the vocabulary conditioned on the latent aspects, $P(w_m|z_k)$ $m \in \{1, ..., M\}$ and $k \in \{1, ..., K\}$, is determined.

Furthermore, the image that is matched with this generated text is represented as some continuous feature vectors. So multivariate Gaussian is introduced into pLSA in order to simulate the generative process of the image. Meanwhile, the proportion of the latent aspects of the image, $P(z|d_i)$, is fixed as that of the text. Specifically, the image feature vectors are sampled from K Gaussian distributions, and each of K Gaussian distributions is corresponding to one of K common latent aspects, z_k $k \in \{1, ..., K\}$. For one fixed latent aspect z_k, the probability density function of the image feature vector f_s is given by the Equation (25).

$$P(f_s|z_k) = \frac{1}{(2\pi)^{C/2} |\varSigma_k|^{1/2}} e^{-\frac{1}{2}(f_s - \mu_k)^{\mathrm{T}} \varSigma_k^{-1} (f_s - \mu_k)} \tag{25}$$

Moreover, the log-likelihood function L_v is given by the Equation (26).

$$L_v = \sum_{i=1}^{D} \sum_{s=1}^{N_i} [\log P(d_i) + \log \sum_{k=1}^{K} P(f_s|z_k) P(z_k|d_i)] \tag{26}$$

When $P(z_k|d_i)$ is fixed, the unobservable parameters $\sum_k$, μ_k can be determined by using EM algorithm to maximize the log-likelihood function L_v

The E-step:

By applying Bayesian formula, we can obtain the Equation (27).

$$P(z_k|d_i, f_s) = \frac{P(f_s|z_k) P(z_k|d_i)}{\sum\limits_{k=1}^{K} P(f_s|z_k) P(z_k|d_i)} \tag{27}$$

The M-step:

The expectation of $E(L_v)$ can be bulit and maximize the expectation of L_v, so we obtain the Equation (28) and (29).

$$\mu_k = \frac{\sum\limits_{i=1}^{D} \sum\limits_{s=1}^{N_i} P(z_k|d_i, f_s) f_s}{\sum\limits_{i=1}^{D} \sum\limits_{j=1}^{N_i} P(z_k|d_i, f_j)} \tag{28}$$

$$\Sigma_k = \frac{\sum\limits_{i=1}^{D} \sum\limits_{s=1}^{N_i} P(z_k|d_i, f_s)(f_s - \mu_k)(f_s - \mu_k)^{\mathrm{T}}}{\sum\limits_{i=1}^{D} \sum\limits_{j=1}^{N_i} P(z_k|d_i, f_j)} \tag{29}$$

When the proportion of the common latent aspects, $P(z|d_i)$, is fixed, the Equation (27), (28) and (29) are alternatively iterated until the convergence condition is satisfied. Thus the parameters $\sum_k$, μ_k are determined.

When a text query arrives, we can transform it into the proportion of the common latent aspects. So we can obtain the proportion of the parameters $\sum_k$, μ_k. Moreover, as for an image in the database, we can calculate its generative probability based on this proportion of $\sum_k$, μ_k. Therefore, we can rank the returned images in terms of their generative probabilities.

4 The Word Classifier based on ELM

The traditional probabilistic semantic models believe that the document contains the potential semantic topics. Therefore, they represent the document as the mixture of the semantic topics without losing the original meanings of the document. The mixture of the semantic topics contains the main semantic meanings of the document. Meanwhile, it has less dimension than the original document. Therefore, the probabilistic semantic models can solve the problem that the original document is too long-winded to operate [1] [2] [4]. At the same time, they need calculate the generative probability of each word conditioned on each semantic topic in the process of transformation from the document into the mixture of the semantic topics. The probabilistic semantic models have the remarkable advantage of the multimodal retrieval because the semantic topics are not restricted by the multimedia modality. So they are easily applied to the multimodal field. Generally speaking, the probabilistic semantic models transform the different types of the multimedia data into the mixture of the semantic topics as their representations. They divide each modality of multimedia document into the smallest elements (textual words or feature vectors). Additionally, they learn the generative probability between the semantic topics of different modalities. What's more they also learn the generative probability of the smallest elements conditioned on the semantic topics. Finally, they calculate the generative probability of the elements of one modality conditioned on the semantic topics of the other modality [24] [25] [34] [35].

However, there are the faults of the probabilistic semantic models. The generative probability of one smallest element conditioned on one semantic topic only statistically take into account the frequency of the occurrence in this semantic topic. So when this semantic topic is picked, a few of elements with larger generative probabilities on this semantic topic are more likely to be sampled. But the multimodal query example itself is neglected in the process of the semantic retrieval. Specifically, the larger generative probability of this element conditioned on this semantic topic only means that this element often occurs in this semantic topic, but it does not necessarily means that this element is related to the query example that contains this semantic topic. In fact, the probabilistic semantic models make the query example more generalized. Therefore, the query example loses its particular quality after it was represented as the mixture of semantic topics. Specifically, the query example is transformed into the generalized semantic environment. Meanwhile it loses the concrete objects. Therefore, the elements with the large generative probability in the transformed semantic topics are picked but they do not necessarily related to the concrete objects of the original query example.

To address this problem, we exploit the Extreme Learning Machine (ELM) to build the multimodal binary classifiers in textual and image modalities. The classifiers determine whether the textual words with the large generative probability are related to the feature vectors or not. Firstly, we make the definition of the sensible word.

Definition 1. (the sensible word). The sensible word is the smallest meaningful element that represent the object or action in the nature.

According to the *Definition 1*, the sensible words can stand by themselves. And the letters are not a sensible word. Although they are smaller than the sensible words, they make no substantial sense by themselves. In this paper, for the sake of convenience, the sensible words only refer to the textual modality, but the sensible words can be applied to any modality, and the smallest meaningful element of any modality can be regarded as the sensible word as *Definition 1.* Furthermore, we assume that the text in the multimodal document can be summarized in some sensible words. The assumption is meaningful because the multimedia is used for recording objects and actions in the nature including human being. So the multimedia including texts can be summarized in some sensible words. Meanwhile we also assume that different modalities of multimedia in the same multimodal document have the same senses [25] [29] [30] [32] [34] [35]. So we can easily conclude *Lemma 1.*

Lemma 1. Every sensible word of the text is mapped to by at least one of the parts of the image in the same multimodal document.

$$\forall w,\ \exists b \textit{ such that } \lambda(b) = w, \tag{30}$$

In the (30) of *Lemma 1* w is a sensible word, and b is one part of images and λ is surjection from b to w.

Proof: The text and the image belong to the same multimodal document. So they have the similar meanings. Every sensible word of the text certainly refer to some parts in the image.

Lemma 1 shows a surjective function that maps parts of the image to sensible words of the text.

Moreover, as the technology of multimedia fast develops, the image feature vectors increasingly exactly represent the whole information of the image.

Lemma 2. Every part of the image is mapped to by at least one image feature vector.

$$\forall b,\ \exists f \ such\ that\ \rho(f) = b, \tag{31}$$

In the (31) of *Lemma 2* b is one part of images, and f is one feature vector, and ρ is surjection from f to b.

Proof: The image feature vectors are the representations of the image in the multimedia and computer vision fields. Therefore, they definitely represent the information of the image. With the development of the technology of the image feature extraction, the image is divided into a few meaningful parts and every part is represented as at least one image feature vector.

Lemma 2 shows a surjective function that maps image feature vectors to parts of the image.

Based on the *Lemma 1* and *Lemma 2*, when the image feature vectors can satisfactorily represent the information of the image we can conclude the *Theorem 1.*

Theorem 1. Every sensible word of the text is mapped to by at least one image feature vector in the same multimodal document.

$$\forall w,\ \exists f \ such\ that\ \mu(f) = w, \tag{32}$$

In the (32) of *Theorem 1* w is one sensible word, and f is one feature vector, and μ is surjection from f to w.

Proof: In the *Lemma 2*, every part of the image is mapped to by at least one image feature vector. In the *Lemma 1*, every sensible word of the text is mapped to by at least one part of the image. So we can conclude that every sensible word is mapped to by at least one image feature vector.

Theorem 1. shows a surjective function that maps image feature vectors to sensible words of the text.

There are some multimodal retrieval methods. The essence of these methods is to use different techniques to learn the surjective function μ in the *Theorem 1.* However, their performance are deteriorated because the input of the surjective function is different modality from the output. It is difficult to build function that maps one modal information to the other one. The method in this paper maps different modal information to the semantic space. It learns the function by means of the semantic layer.

Now we simplify the surjective function. As for each of sensible words, we employ ELM technique to learn the binary function that verify whether this sensible word can be mapped to by feature vectors. Specifically we use the feature vectors of images and the sensible words of their associated texts to train the binary classifiers. The classifiers can determine the binary correlation between the feature vectors and sensible words. And the classifier is used for verifying whether one sensible word is corresponding to any of the feature vectors of the query image or not. Why don't one multi-class classifier is trained to assign the query image into some classes? For one thing, it is inefficient and ineffective to directly use the multi-class classifier to fit the function that maps feature vectors to sensible words. Specifically, the multi-class classifier is more prone to the error than the binary classifier in the multimodal environment. For another, the answer to the query is just a simple *Yes/No* decision for the candidates. The multi-class classifier substitutes for the binary one just as a simple problem is changed into a complex one. Therefore, a simple problem that the classifier makes a *Yes/No* decision for the candidates brings less errors.

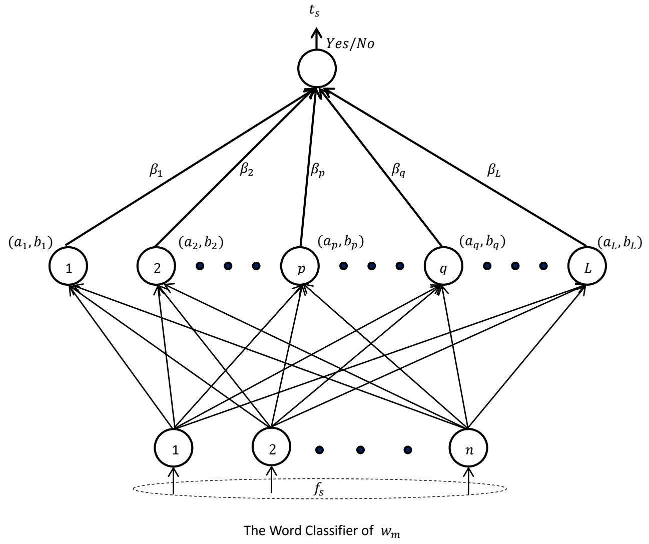

Fig. 1. Structure of The Word Classifier based on ELM

The ELM is an effective and efficient learning algorithms for SLFNs to solve the problem of classification [5] [6] [12]. We assume that all words in the vocabulary are the sensible words. In this paper, the sensible words are called for the words for short. We employ ELM to train the binary classifiers of every word in the vocabulary. We use the word classifier to denote this binary classifier. As for the word classifier of one word w_m where $m = 1, 2, ..., M$, to begin with we build a new training set different from the old training set of our probabilistic semantic model. This new training only is corresponding to the word classifier of w_m. So we need build M new training sets. Each of them is corresponding to one of M word classifiers, where M is the number of words in the vocabulary. As for the new training set of the word classifier of w_m, we collect all of the different feature vectors of all old-training images. We build the samples (f_s, t_s) where $s = 1, 2, ..., N$ and N is the number of the different feature vectors of all old-training images. Among the sample (f_s, t_s), (f_s is the image feature vector and the input sample, and t_s is the output sample. Moreover, if the word w_m is the same with any word of the text associated with the image x_i where $i = 1, 2, ..., D$, every output sample t_s corresponding to every feature vector f_s of the image x_i is set to Yes. Or else every output sample t_s of x_i is No. Furthermore, we exploit the ELM method to train the word classifier of w_m in terms of the corresponding new training set. Specifically, the word classifier of w_m apply ELM to learning the parameters of SLFNs - a_p, b_p and β_p as shown in the Fig. 1. Thus we can obtain the word classifier of this word w_m. Repeat the similar process of building the word classifier for every word in the vocabulary. We can obtain the word classifier of every word of the vocabulary. After we use the multimodal probabilistic semantic model to get the generative probabilities of words, we rank these words in terms of their generative probabilities and we consider them as the candidate words. Additionally, we set a threshold for the word classifiers. As for the ranked candidate words before the threshold, we employ the word classifier of w_u corresponding to the candidate word w_u to verify whether w_u is related to the query image or not, where $u = 1, 2, ..., M$. If the multimodal model does not acquire enough words when it has verified all candidate words before the threshold, it generate words only depend on the generative probabilities from the beginning of the rank. The image feature vectors can be preprocessed without loss of their senses so that they are suitable for classification. If one word w_u where $u = 1, 2, ..., M$ belongs to the text that exist in the same multimodal document with the image x_i, theoretically speaking at least one of the feature vectors of x_i will be classified into the class Yes of w_u by the word classifier of w_u based on $Theorem 2$.

Theorem 2. For any of the sensible words of the text, at least one feature vector of the image in the same multimodal document is classified into its class by its word classifier without taking the error into consideration.

$$\forall w_i,\ \exists f_j \ such\ that\ \mu_i(f_j) = 1 \tag{33}$$

In the (33) of *Theorem 2* w_i is one sensible word, and f_j is one feature vector, and μ_i is the corresponding surjective function of w_i.

Proof: Based on *Theorem 1*, we can draw the conclusion that for any of the sensible words of the text, it can be corresponding to at least one feature vector of the image in the same multimodal document. Meanwhile, there exist the correlation between the feature vectors and the image they belongs to based on *Lemma 2* and the definition of the feature vectors. So there also exist correlation between the corresponging feature vectors and this sensible word based on *Lemma 1* and *Theorem 1*. Therefore, according to the definition of the classifier, when the number of the training image feature vectors is large enough, the classifier can seize the correlation between the feature vectors and the sensible word. So it can filter out the unrelated sensible words that is not correlated with the mapped feature vectors. Thus any image feature vector correlated with this sensible word should be classified into the class of this sensible word by this word classifier. So this word classifier can classify at least one image feature feature vector of the image in the same multimodal document into this sensible word without taking the error into consideration.

Consequently, the word classifier can match the image with its related words and it also can filter out unrelated words. In the addition, the word classifiers can be easily applied to any modality only if the modality follow the *Definition 1*, *Lemma 1* and *Lemma 2*.

5 Experiments

In this section, we study the effectiveness and the efficiency of the multimodal retrieval model based on ELM classifiers. More importantly, we also extend the word classifier to other probabilistic semantic model so as to demonstrate its universality and effectiveness. To begin with, we introduce the experimental environment and provide the definitions of the performance measurements. Furthermore, we measure the performance of searching texts by the image examples based on word classifiers through the classic image annotation. Additionally, we implement the text-based image retrieval to evaluate the performance of the image retrieval by the textual examples. The experimental results demonstrate that this multimodal retrieval model based on ELM classifiers is effective and efficient and what's more the word classifier can be widely applied to other semantic probabilistic models.

5.1 The Experimental Environment and the Performance Measurement

We employ Corel $5K$ dataset to evaluate the performance of the multimodal model. This dataset consists of 5000 images and their associated captions. Among them, 4500 images and captions are regarded as the training set and the rest 500 images and captions are used for the testing set. We directly use the image feature vectors from [33] because the image feature extraction is not the focus of this paper. Moreover, we fit these image vectors for the implementation of the experiments by removing uplicated and irrelevant features in the vectors. At last we keep the rest 12 dimensional feature vectors to implement the experiments. The hardware infrastructure of the experiments is the Intel Core i5-4590 CPU and 8G RAM, and meanwhile the experiments are implemented under the software environment of matlab 2011b and Windows 10.

We employ the performance measurement of HitRate3 of [32] to evaluate the image annotation of the multimodal model and word classifiers, and at the same time we also take into account the time spent on the word classifiers to evaluate the efficiency of word classifiers. HitRate3 is defined as:

HitRate3 (HR3): the average rate of at least one word in the ground truth of a testing image is returned in the top 3 returned words for the testing set[32].

The HR3 describes the accuracy of the image annotation. Obviously, the higher the HR3, the better accuracy of the image annotation.

We use the general precision to evaluate the performance of the text-based image retrieval. And we define the precision as: the average rate of the related images - the related images mean that their ground truth captions contain at least one sensible word of the query text - to the total returned images.

It is obvious that the higher precision, the better the accuracy of the text-based image retrieval. The captions of the images are brief and concise, and therefore we consider every word of all captions as the sensible word. Namely, every word in the vocabulary is the sensible word. In the experiments, the sensible words are called for the words for short.

We set the number of multimodal classifier hidden nodes as 150. Meanwhile, *tribas* function is chosen as active function of ELM.

5.2 Image Annotation

We respectively set the number of the shared latent aspects (topics) to 20, 30, and 40 to implement the experiments. At the same time, we perform both the multimodal retrieval method with word classifiers and that without word classifiers so as to evaluate the effectiveness of the word classifiers. We set a threshold and the word classifiers verify every word whose rank is prior to the threshold. The Fig. 2 shows that the multimodal probabilistic model with word classifiers always has higher accuracy than that without word classifiers. Therefore, the Fig. 2 demonstrates that the word classifiers can improve the accuracy of the multimodal probabilistic semantic model whatever the number of its topics is, because it is able to filter out the unrelated candidate words. On the other hand, the implementing time is so little that the cpu counter considers the time spent on calculating the generative probability of answers nearly instantaneous, when the time for loading the related documents from the hard disk is neglected. Furthermore, we do not take into account the time for loading the related documents and word classifiers into the main memory from the external storage and we assume that these data are stored in the main memory all the time. In the addition, the time for the word classifiers determining whether one word is related to one query image or not is nearly instantaneous too. Therefore, the multimodal retrieval model and the word classifiers also have the excellent efficiency.

Moreover, we adjust the value of the threshold of the word classifiers to study the effect of its change on the accuracy. The Fig. 3 shows that the accuracy of the multimodal model with different numbers of the topics changes as the threshold increases. From the Fig. 3, we can find that the accuracy of the multimodal model rises at first and after it reaches the peak it nearly levels off, because the word classifiers can not find related words from much lower generative probabilities of words. This tendency of the accuracy in the Fig. 3 indirectly demonstrates the validity of the probabilistic semantic model because much lower generative probabilities of words usually are unrelated with the query images. What's more, the improved accuracy also demonstrates the validity of the word classifiers in

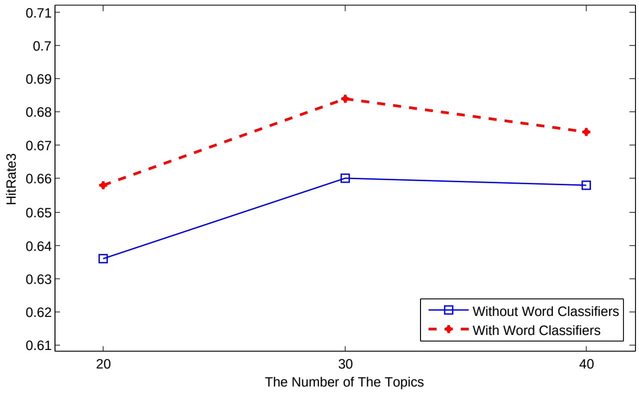

Fig. 2. Accuracy of the Multimodal Retrieval Model

real practice in the Fig. 3. Thus the word classifiers are the effective complement to the probabilistic semantic model.

We compare the multimodal model based on ELM classifiers with traditional semantic models to demonstrate the superiority of our model [32] [40]. Furthermore, we apply the word classifiers to the other semantic models. The improved accuracy of the other models demonstrate that word classifiers are effective. Additionally, the excellent efficiency of the ELM method make implementation of word classifiers nearly instantaneous. Therefore, these experiments demonstrate that ELM are the very effective and efficient method for classification.

Table 1. Performance Comparison with Bayesian Model in Image Annotation

Models	Hit-Rate3
The Multimodal Model	0.636
The Multimodal Model with Word Classifiers	0.658
The Bayesian Model	0.544
The Bayesian Model with Word Classifiers	0.578
The Latent Space Model	0.536
The Latent Space Model with Word Classifiers	0.556
The MLR Based Model	0.486
The MLR Model with Word Classifiers	0.504

The number of the latent aspects of our model and other semantic models are the same 20. The Table 1 shows the accuracy of the models with and without the word classifiers. In the Table 1, the accuracy of our model is higher than that

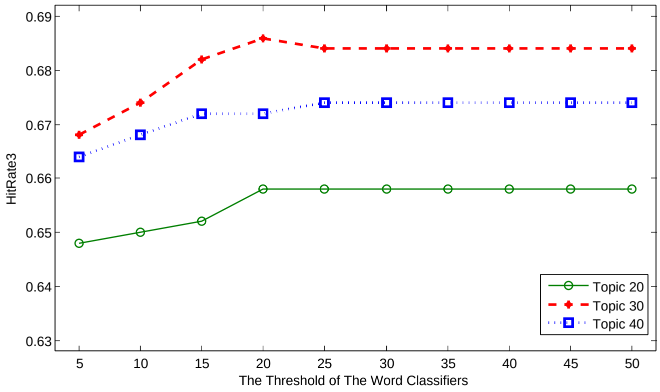

Fig. 3. Accuracy With Changes of Threshold

of other models. More importantly, when we apply the word classifier to other models, the word classifier also can improve their accuracy. The testing time of the word classifiers also is nearly instantaneous owing to the highly efficient ELM method. These experiments can demonstrate that the word classifier can easily be extended to the usual probabilistic semantic models as the effective and efficient complement to them.

The Fig. 4 shows that the accuracy of the Bayesian model changes as the threshold increases. In the Fig. 4, at the beginning the accuracy of the Bayesian model rises and it nearly levels off after it reaches the peak. Both the Fig. 4 and the Fig. 1 demonstrate the expandability and effectiveness of the word classifiers based on ELM.

5.3 Text-based Image Retrieval

In the experiments, we use the precision defined in the above to measure the performance of text-based image retrieval of the multimodal retrieval model with ELM classifiers. If the ground truth caption of the returned image contains at least one word of the query text, this image is regarded as the related image. The precision describes the average rate of the related images to the total returned images. The multimodal text-based image retrieval is a emerging focus compared to the traditional one, so the usual probabilistic semantic models are not designed for it. Therefore, the usual probabilistic semantic models generally have the deteriorated effect on it. The Table 2 shows the precision of the multimodal retrieval model for the text-based image retrieval. In the Table 2, we can see that the word classifiers based on ELM can dramatically improve the precision of the multimodal model although the multimodal model has the unsatisfactory effect on the text-based image retrieval. Therefore, the word clas-

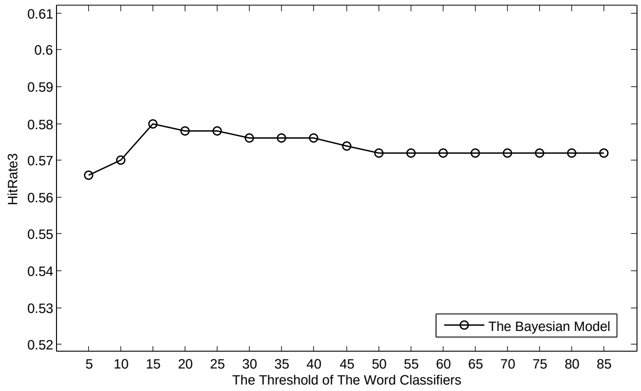

Fig. 4. Effect of Threshold on Bayesian Model

sifiers also have the satisfactory effect on the text-based image retrieval. Owing to the high efficiency of the ELM method, the testing time of the word classifiers is nearly instantaneous.

Table 2. Precision of Text-based Image Retrieval

The Number of The Topics	The Model With Classifiers	The Model Without Classifiers
20 topics	0.403	0.216
30 topics	0.388	0.245
40 topics	0.392	0.218

We compare the multimodal model with word classifiers with the Bayesian model in the Table 2. The multimodal model with word classifiers has the higher precision than the Bayesian model, because the word classifiers improve the performance of the multimodal model.

In short, the word classifiers can be easily extended to the probabilistic semantic models as the effective complement so as to improve the accuracy of this model. Furthermore, owing to the high efficiency of ELM the word classifiers also are efficient. Meanwhile, these experimental results also demonstrate that ELM are the very effective and efficient method for the multimodal classification.

Table 3. Performance Comparison with Bayesian Model in Text-based Image Retrieval

Models	The Precision
The Multimodal Model	0.216
The Multimodal Model with Word Classifiers	0.403
The Bayesian Model	0.242

6 Related Works

The semantic analysis model initially is proposed to handle the notorious semantic gap [19] whose existence in the multimedia document retrieval (MDR) dramatically deteriorates the efficiency of traditional retrieval methods. To tackle the new challenge of the multimodal retrieval, in recent years topic models included by Semantic analysis models are extensively extended into the multimodal field. Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Analysis(pLSA) and Latent Dirichlet Allocation (LDA) are three classical and popular semantic topic models [1] [2] [4]. They were used for reducing the dimensionality of document indexing and meanwhile they can bridge the semantic gap between the human thought and low-level features. Therefore, nowadays they have been widely applied to the multimodal retrieval field.

LSI was first presented by Landauer and P.Foltz in 1998 to reduce the dimension of the documents index [2]. It maps the document into a new semantic space whose dimensionality is far smaller than the original document's index. It has been used for the multimedia and multimodal retrieval [20]. However, there are some intractable flaws in the LSI model. For one thing, LSI is not an incremental model, and therefore the LSI representation of documents need be rebuilt completely once the corpus of documents is changed a little [20]. For another, some parameters of the LSI model intrinsically have not the explicit physical interpretation.

Based on the similar keystone with LSI, Hoffman proposed the Probabilistic Latent Semantic Analysis (pLSA) by introducing the probability theories into the LSI model [1]. This model gives the straightforward physical interpretation to every parameters and probabilistic distributions. At first, pLSA is used for the document clustering [22] [23]. Moreover, pLSA is extended to the computer vision and multimedia fields [36] [37]. In these fields, the image is viewed as a document and it is decomposed into visual words in terms of some methods such as Scaleinvariant feature transform (SIFT) [38]. Recently, pLSA was applied to the multimodal retrieval [34] [35]. In this field, it equally views image words and text ones as the document words to train the pLSA model and it usually introduces extra parameters to balances occurrence frequencies of image words and text ones.

Additionally, like pLSA, LDA is also a generative probabilistic semantic model based on the similar fundamentals [4]. LDA employs the Dirichlet probability to represent the probability distribution of latent semantic topics. It also has been extended to the multimodal field [24] [25].

Some other methods also can effectively address the multimodal retrieval problem [26] [27] [28] [46]. The mathematical matrix and traditional content-based query methods are widely applied to the multimodal retrieval [29] [30]. At the same time, the Principal Component Analysis (PCA) is used for the fusion of the concrete representations of different modalities so as to implement the multimodal retrieval [31].

7 The Conclusion and Future Works

In this paper, we propose a new multimodal retrieval model with the ELM classifiers. The experimental results show that this model are effective and efficient and also demonstrates that the word classifiers based on ELM not only can improve the performance of the probabilistic semantic model but also have the excellent expandability. In the future, we will refine the technique of the word classifiers and apply them to more multimodal models to improve their effectiveness.

8 Acknowledgments

This research is supported by National Natural Science Foundation of China under Grant Nos. 61303016, 61622202 and 61572119; the Fundamental Research Funds of the Central Universities No. N150402005.

9 The References Section

References

1. T. Hofmann: Unsupervised Learning by probabilistic Latent Semantic Analysis. Machine Learning, 42(1-2), 177–196 (2001)
2. a. D. L. T.Landauer, P.Foltz: An Introduction to Latent Semantic Indexing. Discourse Processes. 25, 2-3, 259–284 (1998)
3. K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan: Matching words and pictures. Journal of Machine Learning Research, (3), pp.1107C1135, (2003)
4. D. M. Blei, A. Y. Ng, M. I. Jordan: Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993C-1022 (2003)
5. G.B.Huang, Q.Y.Zhu, C.K.Siew: Extreme Learning Machine: A New Learning Scheme of Feed Forward Neural Networks. In:Proceedings of IEEE International Joint Conference on Neural Networks, pp.985C-990 (2004)
6. G.B.Huang, Q.Y.Zhu, C.K.Siew: Extreme Learning Machine: Theory and Applications. Neurocomputing. 70, 489C-501 (2006)
7. G.B. Huang, D.H.Wang, Y.Lan: Extreme Learning Machines: A Survey. Int. J. Mach. Learn. Cybern. 2(2), 107C-122 (2011)
8. G.B.Huang, L.Chen: Convex Incremental Extreme Learning Machine. Neuro computing. 70, 3056C-3062. (2007)
9. G.B.Huang, L.Chen: Enhanced Random Search based Incremental Extreme Learning Machine. Neurocomputing, 71, 3460C-3468 (2008)

10. G.B. Huang, Y.Q.Chen, H.A.Babri: Classification Ability of Single Hidden Layer Feedforward Neural networks. IEEE Trans. Neural Netw. 11(3), 799-801 (2000)
11. G.B.Huang, Q.Y.Zhu, K.Z.Siew, P.Saratchandran, N.Sundararajan: Can Threshold Networks Be Trained Directly? IEEE Trans. Circuits Syst. 2, 53(3), 187C-191 (2006)
12. G.B. Huang, C.K.Siew: Extreme Learning Machine: RBF Network Case. In: Proceedings of the 8th International Conferenceon Control, Automation, Robotics and Vision, Kunming, China, pp.1029C-1036 (2004)
13. C. Cortes, V.Vapnik: Support Vector Networks, Maching Learning 20(3), 273C-297 (1995)
14. H.J.Rong, G.B.Huang, Y.S.Ong: Extreme Learning Machine for Multi-categories Classificaiton Applications. In: Proceedings of IEEE International Joint Conference on Neural Networks, pp.1709C1713. (2008)
15. G.B.Huang, X.J.Ding, H.Zhou: Optimization Method based Extreme Learning Machine for Classification. Neurocomputing (2010).
16. G.B.Huang, H.Zhou, X.Ding, R.Zhang: Extreme Learning Machine for Regression and Multi-class Classification. IEEE Trans. Systems Man Cybern. B. Cybern. 42(2), 513–529 (2011)
17. G. Wang, Y.Zhao, D.Wang: A Protein Secondary Structure Prediction Framework based on The Extreme Learning Machine. Neurocomputing, 72(1C3), 262C-268 (2008)
18. K. Hornik, Approximation Capabilities of Multilayer Feedforward Networks. Neural Networks. 4, 251C-257 (1991)
19. A.W. M. Smeulders, M.Worring, S. Santini, A. Gupta, R. Jain: Content Based Image Retrieval at The End of The Early Years. IEEE Trans. on Pattern Analysis and Machine Intelligence, 22, 1349C-1380 (2000)
20. Trong-Ton Pham, Nicolas Eric Maillot, Joo-Hwee Lim, Jean-Pierre Chevallet: Latent Semantic Fusion Model for Image Retrieval and Annotation. In: ACM International Conference on Information and Knowledge Management, pp.6–10 (2007)
21. P. Duygulu, K. Barnard, J.F.G. de Freitas, D.A.Forsyth: Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In: European Conference on Computer Vision, pp.97–112 (2002)
22. Y.Mkim, J. F. Pressiot M. R.Amini: An Extension of pLSA for Document Clustering. In: Proceedings of the 17th ACM International Conference on Information and Knowledge Management, pp.1345–1346 (2008)
23. Lingfeng Niu, Yong Shi: Semi-supervised pLSA for Document Clustering. In: IEEE International Conference on Data Mining Workshops, pp.1196–1203 (2010)
24. Duangmanee Putthividhya, Hagai T. Attias, Srikantan S. Nagarajan: Topic regression multi-modal Latent Dirichlet Allocation for image annotation. In: IEEE Computer Vision and Pattern Recognition, (2010).
25. D. M. Blei and M. I. Jordan: Modeling Annotated Data: Proceedings of the 26th ACM Special Interest Group on Information Retrieval (2003)
26. S. L. Feng, R. Manmatha, V. Lavrenko: Multiple Bernoulli Relevance Models For Image And Video annotation. In: Proceedings of The International Conference on Computer Vision and Pattern Recognition, pp.1002–1009 (2004)
27. V. Tseng, J. Su, B. Wang, Y. Lin: Web Image Annotation by Fusing Visual Features and Textual Information. In: Proceedings of ACM Symposium Applied Computing, pp.1056–1060 (2007)
28. J. Li, J. Wang: Automatic Linguistic Indexing of Pictures by A Statistical Modeling Approach. IEEE Transaction on Pattern Analysis and Machine Intelligence 25(19), pp.1075–1088 (2003)

29. Zhuang Yueting, Yang Yi, Wu Fei: Mining Semantic Correlation of Heterogeneous Multimedia Data for Cross-Media Retrieval. IEEE Transactions on Multimedia, 10(2), 221–229 (2008)
30. Yi Yang, Yue-Ting Zhuang, Fei Wu, Yun-He Pan: Harmonizing Hierarchical Manifolds for Multimedia Document Semantics Understanding and Cross-media Retrieval. IEEE Transactions on Multimedia. 10(3), 437–446 (2008)
31. Rasiwasia N., Pereira J.C., Coviello E., Doyle G., Lanckriet G.R.G., Levy R., Vasconcelos N.: A New Approach to Cross-Modal Multimedia Retrieval. In: Proceedings of the 18th International Conference on Multimedia, pp.251C-260 (2010)
32. R. Zhang, Z. M. Zhang, M. Li, W.-Y. Ma, H.-J.Zhang: A Probabilistic Semantic Model for Image Annotation and Multi-modal Image Retrieval. In: IEEE International Conference on Computer Vision, 846–851 (2005)
33. Z. Guo, Z. Zhang, E. P. Xing, and C. Faloutsos: A Max Margin Framework on Image Annotation and Multimodal Image Retrieval. In: Proceedings of the 2007 IEEE International Conference on Multimedia and Expo, pp.504–507 (2007)
34. Pulla Chandrika, C. V. Jawahar: Multi Modal Semantic Indexing for Image Retrieval. In: Proceedings of the ACM International Conference on Image and Video Retrieval, 342–349 (2010)
35. Rainer Lienhart, Stefan Romberg, Eva Horster: Multilayer pLSA for Multimodal Image Retrieval. In: Proceedings of the ACM International Conference on Image and Video Retrieval, pp.1–8. New York (2009)
36. A. Bosch, A. Zisserman, X. Mu noz: Scene Classification via pLSA. In: Proceedings of the European Conference on Computer Vision, 3954, pp.517C-530 (2006)
37. R. Lienhart and M. Slaney: pLSA on Large Scale Image Databases. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 1217C-1220 (2007)
38. D.Lowe: Distinctive Image Feature From Scale-invariant Keypoints. In: Inernational Journal of Computer Vision, 60(2), pp.91–110 (2004)
39. E. Mizutani, S.E. Dreyfus, K. Nishio: On Derivation of MLP Backpropagation From The Kelley-Bryson Optimal-control Gradient Formula and Its Application, In: Proceedings of the IEEE International Joint Confence on Neural Networks, 2, pp.167C-172 (2000).
40. F. Monay, D.G. Perez. On Image Auto-annotation With Latent Space Models. In: the eleventh ACM international conference on Multimedia, 275-278 (2003)
41. G.B. Huang, E Cambria, K Toh, B Widrow: New Trends of Learning in Computational Intelligence. In: IEEE Computational Intelligence Magazine, 10(2), pp:16–17 (2015)
42. D.P. Acharjya, S. Dehuri, S. Sanyal: Computational Intelligence for Big Data Analysis. In: Adaptation, Learning, and Optimization, 19, 87–96 (2015)
43. X. Liu, Y. Dou, J. Yin, L. Wang, En Zhu: Multiple Kernel K-means Clustering With Matrix-induced Regularization. In: 30th AAAI Conference on Artificial Intelligence, 1888-1894, (2016)
44. X. Liu, L. Wang, G.B. Huang, J. Zhang, J. Yin: Multiple Kernel Extreme Learning Machine. In: Neurocomputing, 149, pp:253-264, (2015)
45. X. Liu, L. Wang, J. Yin, Y. Dou, J. Zhang: Absent Multiple Kernel Learning. In:AAAI 2807–2813, (2015)
46. Souja Poria, Iti Chaturvedi, Erik Cambria, Amir Hussain: Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. In: IEEE 16th International Conference on Data Mining, 439–448, (2016)

47. J Cao, K Zhang, M Luo, C Yin, X Lai. Extreme Learning Machine and Adaptive Sparse Representation for Image Classification, Neural Networks, 81, pp:91-102 (2016)
48. J Cao, W Wang, J Wang, R Wang. Excavation Equipment Recognition based on Novel Acoustic Statistical Features. In: IEEE Transactions on Cybernetics, 99, pp:1-13 (2016)
49. J Cao, X Lai, T Chen, J Fan. Accurate and Efficient Scene Recognition With Compact BoW and Ensemble ELM, In: the 12th World Congress on Intelligent Control and Automation, 2058-2063 (2016)