Neurocomputing 274 (2018) 29–36

Contents lists available at ScienceDirect

Neurocomputing

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

Scalable graph based non-negative multi-view embedding for image ranking

Shuhan Qi[a,b], Xuan Wang[a,c,*], Xi Zhang[a,b], Xuemeng Song[d], Zoe L. Jiang[a,c]

[a] Harbin Institute of Technology Shenzhen Graduate School, ShenZhen, Guangdong 518055, China
[b] Shenzhen Applied Technology Engineering Laboratory for Internet Multimedia Application, ShenZhen, Guangdong 518055, China
[c] Public Service Platform of Mobile Internet Application Security Industry, ShenZhen, Guangdong 518055, China
[d] School of Computing, National University of Singapore, 117417 Singapore

ARTICLE INFO

Article history:
Received 7 March 2016
Revised 9 June 2016
Accepted 27 June 2016
Available online 12 April 2017

Keywords:
Image retrieval
Ranking
Multiview embedding

ABSTRACT

Due to the well-known semantic gap, content based image retrieval task is still a challenge problem. The performance of image ranking highly depends on feature representation. In this paper, trying to make a more discriminative feature, we propose a multi-graph based non-negative feature embedding framework for image ranking. In this framework, various image features are embedded into a unified latent space by a learned graph based non-negative multi-view embedding model. In this model, a multi-graph based regularization term, which discovers the intrinsic geometrical and the discriminating structure of the data space, is imposed into the non-negative matrix factorization. The framework learns to find an optimized combination of different Laplacian matrices to approximate the intrinsic manifold automatically. Meanwhile, multiple anchor graphs are utilized to reduce the complexity of computational. Finally, ranking is conducted according to the relevance score inferred by a Markov random field. Extensive experiments prove the effectiveness of proposed method.

1. Introduction

Over the past decades, with the rapid development of internet technology as well as multimedia services, an increasing number of images have been generated and shared on the web like Flicker and Picasa. The fast growing number of web images necessitates effective and efficient image retrieval technologies. As one branch of image retrieval, Content-Based Image Retrieval (CBIR), a technique of retrieving images from large scale image dataset by image query, has been studied extensively [27,40,48,51,53].

Having a better comprehension and representation of query and candidate images should lead to a better retrieval result. The current CBIR methods are usually based on low level visual features (such as texture, color, pixel, etc.), which always fail to describe high-level concepts. There is a famous semantic gap that exists between the low-level image features captured by machine and the high-level visual concepts perceived by human. Due to the semantic gap, the CBIR problem, whose retrieval performance depends on the feature representation, is still one of the most challenging academic problems.

To overcome this problem, some works introduced the multi-model based methods into CBIR [26,43,54]. These methods assumed that using different features which were generated by different visual models have different representation of the underlying data structure. In this way, they tried to bridge the semantic gap by looking for the complementary of various features. Graph-based multi-view manifold ranking (GMMR) framework is a well-known multi-model image retrieval framework [27,40,48,53]. The graph-based ranking model, which ranks data samples with respect to the intrinsic manifold structure, is more meaningful for capturing the semantic relevance degrees. In further, the graph-based multi-view manifold ranking framework builds a intrinsic manifold structure by considering different pair-wise relationships and selecting an optimal combination of manifolds automatically. For the GMMR, some works [27,40] generated a more discriminative and robust representation for queries and candidate images, and some works utilized the manifold ranking model to assign each data sample a relative ranking score directly [48,53]. Both of these methods tried to have a better understanding of queries and

* Corresponding author at: Harbin Institute of Technology Shenzhen Graduate School, ShenZhen, China.
E-mail addresses: shuhanqi@cs.hitsz.edu.cn (S. Qi), wangxuan@cs.hitsz.edu.cn (X. Wang), xizhang@cs.hitsz.edu.cn (X. Zhang), sxmustc@gmail.com (X. Song), zoeljiang@gmail.com (Z.L. Jiang).

http://dx.doi.org/10.1016/j.neucom.2016.06.097
0925-2312/

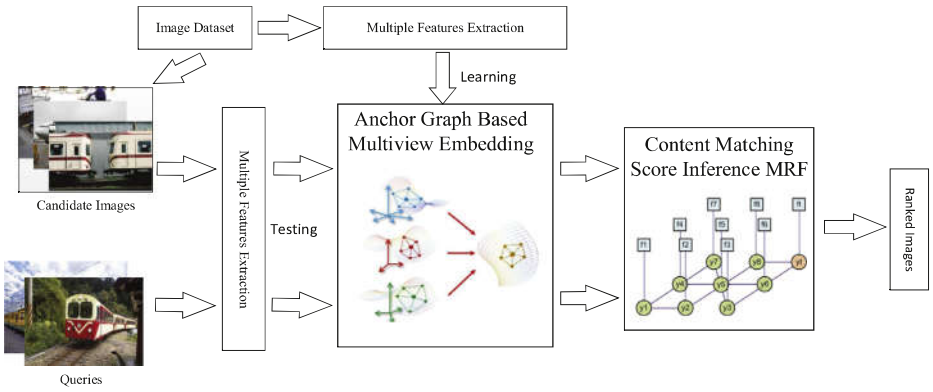

Fig. 1. The framework of the proposed method. In this framework, a learned graph based non-negative multi-view embedding model is utilized to embed multiple image features into a unified latent space and generate the new embedded features about images, then a Markov Random field is constructed by considering the new feature, finally the candidate images are ranked according to the relevance scores.

candidate images by utilizing multiple features that were generated by different models. However, the graph-based multi-view manifold ranking framework has its own drawbacks in interns of handling large scale datasets. It has expensive computational cost, in both graph construction and ranking stage.

In this paper, we propose a ranking method for large scale content based image retrieval. In this method, a graph based non-negative multi-view embedding model is proposed to embed multiple image features into a unified latent space. In this model, a multi-graph based regularization term which discovers the intrinsic geometrical of the data space is imposed for reinforcing the non-negative matrix factorization. By providing the graph Laplacians of various features, the framework learns to find an optimal combination of these Laplacians to approximate the ideal intrinsic manifold. Further, to make the embedding model more efficient, a scalable anchor graph is introduced. Finally, images are ranked according to the relevance scores inferred by a Markov random field. The Fig. 1 shows the illustration of the proposed framework. Extensive experiments are done to prove the effectiveness of proposed method.

The contribution of this paper is in two folds: First, a multi-view embedding based image ranking framework is proposed. In this framework, a multi-graph based multi-view non-negative embedding model is obtained by unsupervised learning. The multi-view embedding model is able to map the multiple image features into a optimal unified latent space. Second, to solve the problems of large storage requirement and extensive computation in multi-view embedding model, an anchor graph based efficient graph construction method is proposed.

2. Related work

In this section, we briefly introduce the related work on image ranking and graph based matrix factorization respectively, which are highly related to our work.

2.1. Query and image ranking

Query prediction is a meaningful problem in information retrieval. One of challenges in CBIR is to convert a textual query into an amenable form for visual search. Image annotation and labeling tasks reverse this problem and tag images with key words that can be used for retrieval [21,41,46]. Recently, the problem of complex query has been widely studied. Many of the improvements showed stem from exploiting query term re-weighting [2–4,23] and query reduction [1,21,22] approaches. Bendersky and Croft [2] developed a technique that assigned weights to identify key concepts in the verbose query, which had been observed to improve the retrieval effectiveness. In [31], a heterogeneous probabilistic network framework was proposed. In the framework, the authors integrated three layers of relationships, i.e., the semantic-level, cross-modality level as well as visual-level. These mutually reinforced layers were established among the complex query and its involved visual concepts, by harnessing the contents of images and their associated textual cues. Kumaran and Allan [21,22] proposed an interactive query induction approach, which presented the users with the top 10 sub-queries along with corresponding top ranking snippets. In [33], a query-adaptive graph-based learning approach was proposed to estimate the images relevance probabilities. This method was evaluated by three applications, namely, image meta search, multilingual image search, and Boolean image search. Recent work [11] proposed a fast democratic aggregation and query fusion method, which embedded weak spatial context in the kernel construction to depress co-occurrence caused by local feature detector.

Graph-based methods performed well in image re-ranking [6,18,21], like emotional image analysis, video annotation and 3D object retrieval. In [45,52], the authors investigated the performance of different features on different kinds of images, and adopted a multi-graph learning framework to solve the image retrieval problem. In [10], the authors proposed a method that constructed multiple hypergraphs for a set of 3-D objects based on their 2-D views and then used these hypergraphs to recognize and retrieval the 3-D objects. In [50], Yang et al. discussed that such graph-based methods are sensitive to the bandwidth parameter of Laplacian matrix. In work [49], Yang et al. proposed an Local Regression and Global Alignment (LRGA) based semi-supervised learning method for image retrieval. In this model, the Laplacian matrix were learned by a local linear regression instead of calculating the pair-wise distances in the whole dataset. In [47], the authors proposed the Bregman divergence to solve Co-Ranking problem(CoR) by leveraging fruitful information from manual semantic labeling (i.e.,tags) and associated images, which leaded to the technique of co-ranking images and tags, then a representative method that aimed to explore the reinforcing relationship between image and tag graphs was introduced. Nie et al. [32,34] proposed a scheme that was able to enrich textual answers in Question and Answer (QA) with appropriate media data. Unfortunately, such a graph-based approach has very high computational complexity, due to the computation of the distance between all image pairs and the computation of the pseudo-inverse of adjacency matrix. Frequent pattern mining was used for removing outliers in [35]. Each image was described as a transaction (or pattern). A pattern consisted of items which were the visual words located on images

interest points. Voravuthikunchai et al. [39] proposed frequently closed patterns which gave excellent re-ranking results. In [12], the authors gave graph mining techniques enriched queries by identifying query concepts and adding relevant synonyms as well as semantically related terms.

2.2. Graph based matrix factorization

The Non-negative Matrix Factorization (NMF) method was proposed to learn a low-rank representation of objects like human faces and text documents [24,25]. Meanwhile the NMF has attracted considerable attention for learning the effective representation of images. However, NMF performs this learning in the Euclidean space. It fails to discover the intrinsic geometrical and the discriminative structure of the data space. Jin et al. [17] proposed a low-rank matrix factorization, in which a manifold regularization term was added to the TSVD framework to leverage regularization term and matrix factorization. Cai et al. extended the transitional NMF to graph regularized Non-negative Matrix Factorization (GrNMF) in [5] to avoid this limitation by incorporating a geometry-based regularizer. Wang et al. [42] proposed an new unified feature selection and graph regularization algorithm, namely Adapt GrNMF. Guo et al. [13] provided an novel method of Robust Non-negative Matrix Factorization with discriminate ability (RNMFD) to tackle several problems, i.e. sensitivity to noise data, trivial solution problems, and ignoring the discriminative information. Lin and Pang [27] learned a sparse representation and proposed a method called Graph Regularized Non-negative Matrix Factorization with Sparse Coding (GRNMF_SC). Tao et al. [38] utilized multiple graph integration for low rank matrix approximation to boost the low decomposition performance caused by graph selection.

The NMF is actually an unsupervised method without making use of prior information of data, He et al. [14] not only utilized the local structure of the data by graph Laplacian, but also incorporated pairwise constraints generated among labeled data into NMF framework. Sun et al. [36] proposed a novel matrix decomposition algorithm, called Graph regularized and Sparse Non-negative Matrix Factorization with hard Constraints (GSNMFC), which incorporated a graph regularizer and hard prior label information as well as sparseness constraints as additional conditions to uncover the intrinsic geometrical and discriminative structures of the data space. Lu et al. [30] extended the recently proposed low-rank matrix with manifold regularization (MMF) method and adaptive graph regularizer (LMFAGR), which simultaneously seeked graph weight matrix and low-dimensional representations of data and incorporated both of them into an unified framework. The standard NMF adopts a least square error function as the empirical likelihood term in the model, which is sensitive to the noise and outliers, Feng et al. [9] proposed an noise robust NMF method named as Locally Weighted Sparse Graph regularized Non-negative Matrix Factorization (LWSG_NMF) by reformulating the empirical likelihood term of the standard NMF and imposing a sparse noise term explicitly.

3. Graph based non-negative multiview embedding for ranking

3.1. Multiple graph based non-negative embedding

Given N images and their feature vector set $X = \{X_n\}, n = 1, \ldots, N$. For each image, k different types of features are extracted. These k features are further concatenated into an vector as the feature of the image, and thus it is $x_i = [x_i^{(1)}, \ldots, x_i^{(k)}] \in R_+^{D\times 1}$, where $x_i^{(k)}$ is the kth feature of data. The feature vectors of N images are organized as a non-negative matrix $X = [X_1, \ldots, X_N] \in R_+^{D\times N}$, where the nth column x_n of X is the feature vector of the nth data point. The aim of an non-negative embedding model is to locate two non-negative matrices B and H whose product approximates well the original matrix as:

$$\arg\min_{B,H} \| X - BH\|^2, \tag{1}$$

where $B \in R_+^{D\times P}$ can be regarded as a set of basis vectors, and $H \in R_+^{P\times N}$ can be regarded as the new representation of images that coding with respects to the basis B. In this way, features with different modalities are encoded into a new feature space by the non-negative embedding model.

However, the aforementioned model has some drawbacks. At first, the concatenation of different features deals with all the views equally and ignores the correlation of different features; second, it fails to discover the intrinsic geometrical and the discriminating structure of data space, which is essential for image representation.

In this paper, to overcome the drawbacks, the following graph regularization term can be added into the embedding model:

$$\begin{aligned} O(H) &= \frac{1}{2}\sum_{n,m=1}^{N} \| h_n - h_m\|^2 W_{nm} \\ &= Tr\left(HDH^T\right) - Tr\left(HWH^T\right) \\ &= Tr\left(HLH^T\right), \end{aligned} \tag{2}$$

where h is a column of H, $W \in R^{N\times N}$ is an affinity matrix, $D \in R^{N\times N}$ is a diagonal matrix, the entries of which are column sums of W, i.e., $D_{mm} = \sum_{n=1}^{N} W_{nm}$, and $L = D - W$ is the graph Laplacian matrix. The graph regularization term can measure the smoothness of feature representation in H. By minimizing this regularization term, two feature vectors h_n and h_m are expected close to each other if the original feature x_n and x_m are close (i.e., W_{nm} is big). By using the graph regularizer, intrinsic geometry information of data distribution can be imposed.

For K different features, there might be K different graph Laplacian candidates $[L_1, \ldots, L_K]$. Assuming that the ideal hidden geometric structure can be explicated by an optimal linear combination space of these initial different manifolds. Theoretically, the integration of multiple manifold structures can be formulated as

$$\begin{aligned} &L = \sum_{g=1}^{K} \gamma_g L_g \\ &s.t.\ \sum_{g=1}^{K} \gamma_g = 1, \gamma_i \geq 0, \end{aligned} \tag{3}$$

where γ_g is the weight of graph Laplacian L_g. For each feature, the heat kernel weighting is utilized to build the affinity matrix W^g:

$$W_{nm}^g = \begin{cases} e^{-(\|x_n^g - x_m^g\|^2)/\sigma}, & \text{if } (n,m) \in \varepsilon, \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where x^g is an specific type of feature and σ is a predefined constant. It is worth noting that for each data point x_j, we find its ε nearest neighbors and make connection between x_j and these neighbors in affinity matrix W.

Combining the multiple graph based regularizer with the original non-negative embedding model, the loss function is formed as

$$\begin{aligned} O(B,H,\gamma) &= ||X - BH||^2 + \alpha \sum_{g=1}^{K} \gamma_g Tr(HL_gH^T) + \beta||\gamma||^2 \\ &= Tr(X^TX) - 2Tr(X^TBH) + Tr(H^TB^TBH) \end{aligned}$$

$$+\alpha\sum_{g=1}^{K}\gamma_g Tr(HL_gH^T)+\beta||\gamma||^2$$
$$s.t.\ B\geq 0, H\geq 0, \sum_{g=1}^{K}\gamma_g=1, \gamma_g\geq 0. \tag{5}$$

The L2 norm regularization term $||\gamma||^2$ is added into the model to avoid the overfitting to a single graph.

3.2. Optimization

The objective function is not convex if both the matrices B and H are needed to be optimized. Here we solve the problem by initializing (B, H) as the non-negative matrix factorization of X, then optimizing (B, H) and γ alternatively.

On optimizing B and H, by fixing γ, the problem can be transformed as follows:

$$\begin{aligned}&\arg\min_{B,H}||X-BH||^2+\alpha\sum_{g=1}^{K}\gamma_g Tr\left(HL_gH^T\right)\\ &s.t.\ B\geq 0, H\geq 0, \sum_{g=1}^{K}\gamma_g=1, \gamma_g\geq 0.\end{aligned} \tag{6}$$

We deploy an Lagrange multiplier for constraint condition and then use the KKT condition to solve the problem [5]. The above equation leads to the following updating rules:

$$\begin{aligned}b_{t+1}&=\frac{XH^T}{BHH^T}b_t,\\ h_{t+1}&=\frac{B^TX+\alpha\sum\gamma_g HW_g}{B^TBH+\alpha\sum\gamma_g HD_g}h_t.\end{aligned} \tag{7}$$

The optimization of γ is solved by fixing (B, H), and the problem is transformed into:

$$\begin{aligned}&\arg\min_{\gamma}\sum_{g=1}^{K}\gamma_g Tr\left(HL_gH^T\right)+(\beta/\alpha)||\gamma||^2\\ &s.t.\ \sum_{g=1}^{K}\gamma_g=1, \gamma_g\geq 0.\end{aligned} \tag{8}$$

This is a constrained quadric programming problem, and can be easily solved by a quadric optimization solver.

3.3. Scalable graph construction

The aforementioned model has two limitations, the first one is the strategy of building the graph, which needs to search all the pair-wise relationships in the whole dataset, the second one is the requirement of solving the matrix multiplications $\Sigma\gamma_g HW_g$ in optimization. Both of them are time-consuming and have a large storage requirement.

Inspired by work [29,44], we propose an anchor graph model to overcome the shortcomings in two perspectives: scalable graph construction and efficient embedding computation.

Here we introduce how to utilize anchor graph to model the data. Given a dataset $X=[x_1,\ldots,x_N]\in R_+^{D\times N}$ with N samples in D dimension, we aim to obtain a set of representative anchors $U=[u_1,\ldots,u_N]\in R_+^{D\times K}$ in the same feature space with original samples. Then each samples in the manifold can be locally approximated by a linear combination of its neighbor anchors:

$$\begin{aligned}&\arg\min_{z_i}\ \|x_i-Uz_i\|^2\\ &s.t.\sum_{j=1}^{K}z_{ji}=1, z_{ji}\geq 0,\end{aligned} \tag{9}$$

where $Z=[z_1,\ldots,z_N]\in R_+^{K\times N}$ is a weight matrix that measures the potential relationships between the data samples and the anchors. Meanwhile, the data samples in the original feature space are mapped into a new space where U can be seen as a set of basis and Z is the K dimensional representations of the data samples.

In practice, the anchors are usually selected by a clustering method such as K-means and local weights Z are defined by

$$z_{ji}=\frac{\exp\left(-t^2(x_i,u_j)/\lambda\right)}{\sum_{l=1}^{K}\exp\left(-t^2(x_i,u_l)/\lambda\right)}, \tag{10}$$

where $t(\cdot,\cdot)$ is a distance function, λ is the bandwidth parameter and K is the number of anchor points that $K\ll N$.

Based on the local weights Z, the adjacency matrix between data samples can be derived

$$W=Z^TZ. \tag{11}$$

From Eq. (11), we can see that if two samples are correlative ($w_{ij}>0$), they should share at least one common anchor point, otherwise $w_{ij}=0$. Since the matrix Z is non-negative and highly sparse, the matrix W is also a positive semi-definite and sparse matrix, which is consistent with fact that most of the points in the graph only have limit number of neighbors.

Compared with the graph construction in Section 3.1, the anchor graph only needs to build the pair-wise relationships between samples and anchors. So the construction has a complexity in $O(NK)$, since $K\ll N$, the construction is linear to the dataset.

When solving the matrix multiplication, the problem can be converted into $\Sigma\gamma_g HZ_g{}^TZ_g$ with complexity in $O(PN+KN)$, which is also linear to the dataset.

3.4. Deep learning features

Recent years have witnessed an important breakthrough in machine learning methods, which are known as deep learning. The deep learning methods include a family of machine learning algorithms that attempting model high-level abstractions in data by employing deep architectures composed of multiple non-linear transformations [31]. Recently deep learning techniques have achieved some success in computer vision and other applications [20,28,37]. In work [7], the experiments suggest that the deep models that are fully supervised trained on a fixed large scale image dataset can be re-purposed to novel generic task. In this paper, we try to utilize two deep learning models that are trained on the ImageNet dataset[1] to generate image features for feature embedding.

The AlexNet model. The first model we utilize here is the Alexnet model [20] from CAFFE [16]. The model is trained on the ImageNet ILSVRC-12, which contains more than 1 million images that belongs to 1000 categories. The Alexnet contains 8 learned layers, which are composed by 5 convolutional layers and 3 fully-connected layers, together with several ReLU activation and max pooling layers. In this paper, the output of the last fully connected layer with 1000 dimension is extracted as image representation. The L2-norm and Signed Square Root are applied for feature normalization.

The Network in Network (NIN) model. The second model we utilize is the NIN-Imagenet model [28] from CAFFE. Different from the conventional convolutional layers, which uses linear filters followed by a non-linear activation function to scan the input, the NIN builds micro neural networks, which are named inception, with more complex structures to abstract the data within the receptive field. By using inception, people can use much less parameter to build a more complex deep learning model. The NIN model

[1] http://www.image-net.org/ .

contains 12 learning convolutional layers, as well as several ReLU activation and pooling layers. The output of last average pooling layer with 1000 dimension is taken as image representation. The L2-norm and Signed Square Root are applied for feature normalization.

3.5. Matching score inference

In this paper, we employ a Markov Random Field (MRF) [19] model to infer the matching scores of candidate images to the query.

The configuration of MRF is defined as below. The latent variable y is the matching score of an image, which stands for how relevance the candidate image is to the query. The single site potential is defined as:

$$\psi_i(y_i) = \frac{1}{1+t}, \tag{12}$$

where t is the initial score of the image, $t = e^{\beta}$ for the query while $t = e^{-\beta}$ for candidate images, and β is an empirically chosen constant.

The pairwise potential of MRF is defined as:

$$\psi_{i,j}(y_i, y_j) = |y_i - y_j| s_{i,j}, \tag{13}$$

where $s_{i,j}$ is the similarity between two images. The similarity $s_{i,j}$ between two images is defined as

$$s_{i,j} = exp\left(-\frac{\| h_i - h_j \|_2^2}{2\sigma^2}\right), \tag{14}$$

where h_i and h_j are the unified features of images after multiview embedding, σ is empirically chosen as the mean distance of all data. The energy function is defined as follows:

$$\epsilon(y) = \sum_i \psi_i(y_i) + \sum_i \sum_j \psi_{i,j}(y_i, y_j), \tag{15}$$

and we aim to maximize the following formulation:

$$p(y|M) = \frac{1}{Z} \exp(-\epsilon(y)), \tag{16}$$

where Z is the normalization parameter.

The loop belief propagation method [15] is employed to optimize the objective function. After optimization, the belief of matching score is obtained. The retrieval results are ranked according to the matching scores.

4. Experiments

4.1. Experimental settings

Experiments were conducted on the *a-Pascal* dataset [8]. The *a-Pascal* dataset contains 12,965 images in variety of natural poses, viewpoints and orientations. The image dataset consists of 6340 training images and 6335 testing images. Each image belongs to at least one of 20 semantic classes such as *people, bird, cat, cow, etc.*

To evaluate the performance of our framework, we conducted our image ranking experiments on the testing images set of *a-Pascal*. As some of classes have few related images, for better comparison, we only chose those classes with more than 100 related images. We thus selected the 16 classes: *aeroplane, bicycle, bird, boat, bottle, car, cat, chair, diningtable, dog, horse, motorbike, person, pottedplant, sofa* and *tvmonitor*. For each class, we conducted the image retrieval on the testing dataset. By deploying our framework on the testing dataset, we first computed the relevance scores of testing images and queries, and then utilized the relevance scores for ranking. In the image retrieval task, the queries for retrieval are important. In this paper, the query images were generated as the positive images of each semantic class.

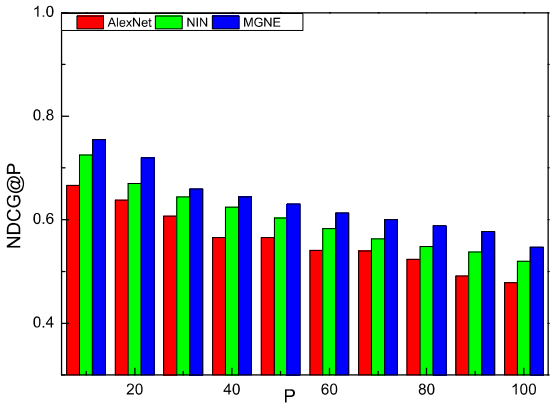

Fig. 2. The average NDCG performance of comparative method in P from 10 to 100.

We conducted the experiments of using our multi-graph based Non-negative embedding (MGNE) to retrieve image. We also conducted the experiments of retrieve by using AlexNet and Network in Network (NIN) feature individually as comparison experiments.

4.2. Evaluation measures

Generally, the image retrieval results are displayed screen by screen. Too many images in a screen may confuse the users and drop the experience evidently. Images in the top pages attract the most interests and attentions from users. Therefore, the precision at P metric is significant to evaluate the image retrieval performance. In this paper normalized discounted cumulative gain (NDCG) is used to evaluate the performance of different methods. NDCG is a standard measure for evaluating ranking algorithms. NDCG of the first P images in a ranked image list is defined as below:

$$NDCG_P = \frac{DCG_P}{IDCG_P}, \tag{17}$$

where

$$DCG_P = \sum_{i=1}^{P} \frac{2^{re_i} - 1}{\lg(i+1)} \tag{18}$$

$$IDCG_P = \sum_{i=1}^{P} \frac{2^{1} - 1}{\lg(i+1)}. \tag{19}$$

In this paper, $re_i \in \{0, 1\}$ denotes that whether the candidate image is related to the query, $re_i = 1$ means that the image and the query is relevant, and $re_i = 0$ means that the image and the query is irrelevant. The first 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 images were evaluated by NDCG in our experiments.

4.3. Experimental results

In this experiment, we set each training image from the selected 16 classes as the query image and retrieved the related images in the testing dataset. The retrieval performance was averaged.

The experimental results of all the comparison methods are shown in the Fig. 2. In our experiments, we set the number of anchor points as 1500. In the later part we will show the impact of the number of anchor point to the experiment performance. In Fig. 2, we can see that our proposed multi-graph based non-negative embedding method achieves the best performance over three methods for comparison. The performance of proposed

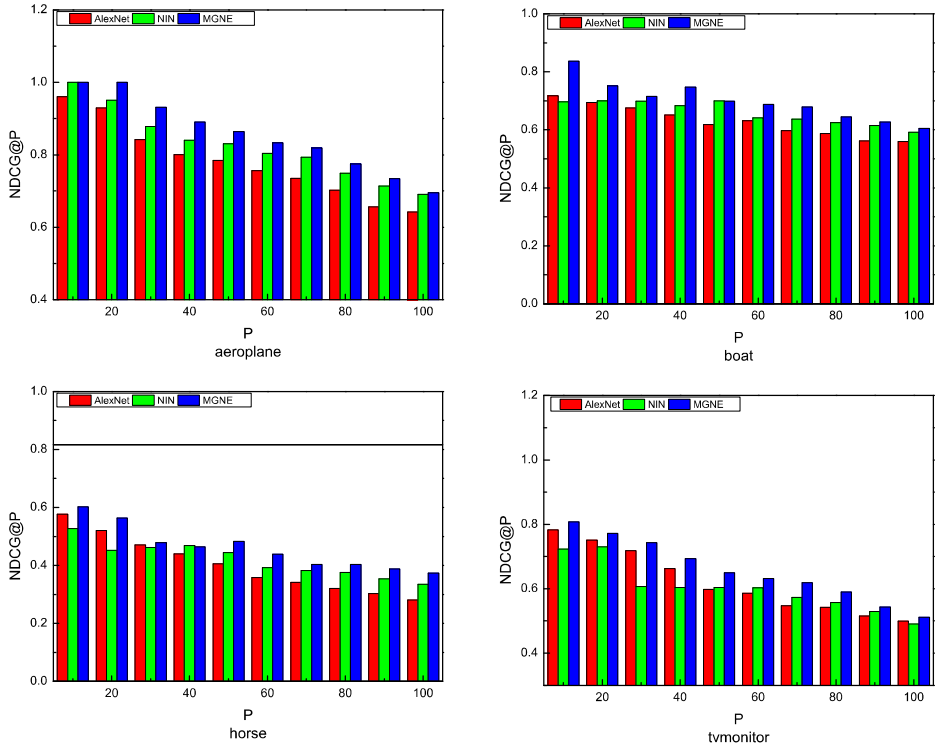

Fig. 3. Comparative ranking results for aeroplane, boat, horse, tvmonitor.

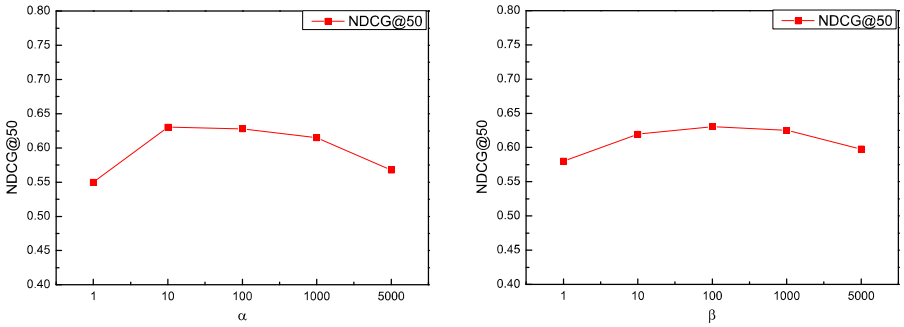

Fig. 4. The influence of versus tradeoff parameters α and β.

method is 0.75 at $P = 10$ and drops to 0.55 at $P = 100$, which outperforms the AlexNet and NIN method by 12.3% and 5.3% respectively. The second best result is NIN, whose performance is 0.72 at $P = 10$ and drops to 0.51 at $P = 100$. The experimental results may suggest that although the deep learning can generate a good representation for images, the semantic gap still exists to affect the performance. By using the proposed method, features generated by different deep learning models can be fused together into a unified latent space with a better discriminative ability. In this way, the proposed method leads to a better image ranking performance. In Fig. 3, we also show some example results on semantic classes of aeroplane, boat, horse and tvmonitor. In these example results, we can also see some similar trends with Fig. 2.

4.4. Parameters setting

The tradeoff parameters of α and β have directly affects on the image representation and retrieval performance. We conducted experiments to test the sensitivity of the performance to these parameters. The left plot of Fig. 4 shows performance of versus α, which varies within the set {1, 10, 100, 1000, 5000}. When the value of α less than 10, the performance is worse. This is because the parameter α decides the importance of graph Laplacian, and in the situation of $\alpha = 0$ the model become a normal NMF model. As the value of α increases, the performance grows and trends to be much more stable until $\alpha = 5000$. The performance decreases when $\alpha = 5000$ may because the weight of graph Laplacian is too large and the model overfits to the graph regularization. The right plot of Fig. 4 shows performance of versus β that varies within the set {1, 10, 100, 1000, 5000}. The effect of β is to increase the diversity of graph regularization and prevent the graph Laplacian from overfitting to a single graph. When $\beta < 10$, the performance is worse. As the value of β is larger than 10 the performance becomes stable and makes a peak when $\beta = 100$.

In Fig. 5, we also show the impact of different anchor points' amount on the image ranking performance. The baseline method is the graph based non-negative embedding method that generates the affinity matrix directly instead of using the anchor graph. In this experiment, we can see that the performance is very sensitive to the anchor points' amount, when the number of anchor points is less than 600. When the number of anchor points exceeds 600, the growth of performance becomes slower and the performance approximates to the baseline gradually. It suggests that 600 anchor points are enough to simulate the manifold structure of data distribution. As the number of anchor points increases to 1500, the proposed method slightly outperforms the baseline. It might be due to

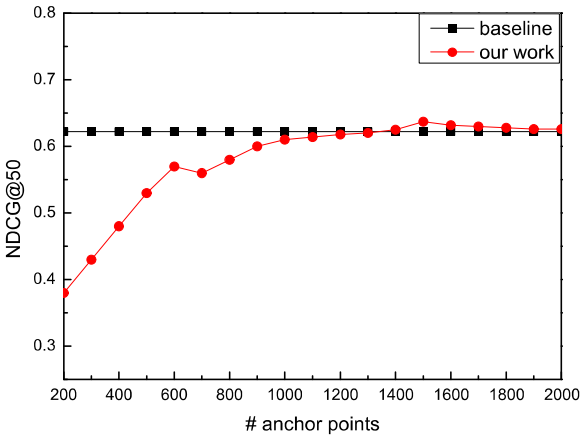

Fig. 5. Performance versus different number of anchor points.

the fact that the anchor graph, using a subset of data points to represent the whole dataset, which reduces the noise in the dataset. In this way, the generated feature becomes more robust.

5. Conclusion

In this work, we presented a large scale content based image ranking framework. In this model, multiple image features were extracted by Alexnet and NIN model, respectively. Then the image features were embedded into a unified latent space by an learned multi-graph based non-negative multi-feature embedding model. Meanwhile, multiple anchor graphs were utilized to reduce the complexity of computational. Finally, a Markov random field was constructed by the query and the testing data, and results were ranked according to the relevance scores, which were inferred by loopy belief propagation. We verified the effectiveness of the proposed method by extensive experiments.

Acknowledgments

This work is supported by Basic Research Project of Shenzhen, China, (No. JCYJ20140417173156099), by International Exchange and Cooperation Foundation of Shenzhen City, China (No. GJHZ20150312114149569), by Science and Technology Planning Project of Guangdong Province (2016A040403046), by Shenzhen Applied Technology Engineering Laboratory for Internet Multimedia Application of Shenzhen Development and Reform Commission (No. [2012]720), by Public Service Platform of Mobile Internet Application Security Industry of Shenzhen Development and Reform Commission (No. [2012]900).

References

[1] N. Balasubramanian, G. Kumaran, V.R. Carvalho, Exploring reductions for long web queries, in: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2010, pp. 571–578.
[2] M. Bendersky, W.B. Croft, Discovering key concepts in verbose queries, in: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2008, pp. 491–498.
[3] M. Bendersky, D. Metzler, W.B. Croft, Learning concept importance using a weighted dependence model, in: Proceedings of the ACM International Conference on Web search and Data Mining, ACM, 2010, pp. 31–40.
[4] M. Bendersky, D. Metzler, W.B. Croft, Parameterized concept weighting in verbose queries, in: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2011, pp. 605–614.
[5] D. Cai, X. He, J. Han, T.S. Huang, Graph regularized nonnegative matrix factorization for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1548–1560.
[6] D. Delgado, J. Magalhaes, N. Correia, Assisted news reading with automated illustration, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2010, pp. 1647–1650.
[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition, in: Proceeding of the International Conference on Machine Learning, 32, 2014, pp. 647–655.
[8] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1778–1785.
[9] Y. Feng, J. Xiao, K. Zhou, Y. Zhuang, A locally weighted sparse graph regularized non-negative matrix factorization method, Neurocomputing 169 (2015) 68–76.
[10] Y. Gao, M. Wang, D. Tao, R. Ji, Q. Dai, 3-d object retrieval and recognition with hypergraph analysis, IEEE Trans. Image Process. 21 (9) (2012) 4290–4303.
[11] Z. Gao, J. Xue, W. Zhou, S. Pang, Q. Tian, Fast democratic aggregation and query fusion for image search, in: Proceedings of the ACM International Conference on Multimedia Retrieval, ACM, 2015, pp. 35–42.
[12] J. Guisado-Gámez, D. Dominguez-Sal, J.L. Larriba-Pey, Massive query expansion by exploiting graph knowledge bases for image retrieval, in: Proceedings of the International Conference on Multimedia Retrieval, ACM, 2014, p. 33.
[13] Y. Guo, G. Ding, J. Zhou, Robust nonnegative matrix factorization with discriminability for image representation, in: Proceedings of the IEEE International Conference on Multimedia and Expo, IEEE, 2015, pp. 1–6.
[14] Y.C. He, H.T. Lu, L. Huang, X.H. Shi, Non-negative matrix factorization with pairwise constraints and graph laplacian, Neural Process. Lett. 42 (1) (2015) 167–185.
[15] A.T. Ihler, J. Iii, A.S. Willsky, Loopy belief propagation: convergence and effects of message errors, J. Mach. Learn. Res. 6 (2005) 905–936.
[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2014, pp. 675–678.
[17] T. Jin, J. Yu, J. You, K. Zeng, C. Li, Z. Yu, Low-rank matrix factorization with multiple hypergraph regularizer, Pattern Recognit. 48 (3) (2015) 1011–1022.
[18] P. Jing, Z. Ji, Y. Yu, Z. Zhang, Visual search reranking with relevant local discriminant analysis, Neurocomputing 173 (2016) 172–180.
[19] R. Kindermann, J.L. Snell, et al., Markov Random Fields and Their Applications, 1, American Mathematical Society Providence, 1980.
[20] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[21] G. Kumaran, J. Allan, A case for shorter queries, and helping users create them, in: Proceedings of the International Conference of the North American Chapter of the Association for Computational Linguistics C Human Language Technologies, 2007, pp. 220–227.
[22] G. Kumaran, J. Allan, Effective and efficient user interaction for long queries, in: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2008, pp. 11–18.
[23] M. Lease, J. Allan, W.B. Croft, Regression rank: learning to meet the opportunity of descriptive queries, in: Advances in Information Retrieval, Springer, 2009, pp. 90–101.
[24] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.
[25] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in: Proceedings of the Advances in Neural Information Processing Systems, 2001, pp. 556–562.
[26] J. Li, Y. Wu, J. Zhao, K. Lu, Multi-manifold sparse graph embedding for multi–modal image classification, Neurocomputing 173 (2016) 501–510.
[27] C. Lin, M. Pang, Graph regularized nonnegative matrix factorization with sparse coding, Mathematical Problems in Engineering, 2015, Hindawi Publishing Corporation, 2015.
[28] M. Lin, Q. Chen, S. Yan, Network in network, in: Proceeding of International Conference on Learning Representations, 2014.
[29] W. Liu, J. He, S.F. Chang, Large graph construction for scalable semi-supervised learning, in: Proceedings of the International Conference on Machine Learning, 2010, pp. 679–686.
[30] G.F. Lu, Y. Wang, J. Zou, Low-rank matrix factorization with adaptive graph regularizer, IEEE Trans. Image Process. 25 (5) (2016) 2196–2205.
[31] Y. Bengio, A.C. Courville, P. Vincent, Unsupervised feature learning and deep learning: A review and new perspectives 1 (2012) CoRR, abs/1206.5538.
[32] L. Nie, M. Wang, Y. Gao, Z.J. Zha, T.S. Chua, Beyond text QA: multimedia answer generation by harvesting web information, IEEE Trans. Multimed. 15 (2) (2013) 426–441.
[33] L. Nie, M. Wang, Z.J. Zha, T.S. Chua, Oracle in image search: a content-based approach to performance prediction, ACM Trans. Inf. Syst. 30 (2) (2012) 13.
[34] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, T.S. Chua, Disease inference from health-related questions via sparse deep learning, IEEE Trans. Knowl. Data Eng. 27 (8) (2015) 2107–2119.
[35] S. Nowozin, K. Tsuda, T. Uno, T. Kudo, k. Bakir, Weighted substructure mining for image analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.
[36] F. Sun, M. Xu, X. Hu, X. Jiang, Graph regularized and sparse nonnegative matrix factorization with hard constraints for data representation, Neurocomputing 173 (2016) 233–244.
[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2015, pp. 1–9.
[38] L. Tao, H.H. Ip, Y. Wang, X. Shu, Low rank approximation with sparse integration of multiple manifolds for data representation, Appl. Intell. 42 (3) (2015) 430–446.

[39] W. Voravuthikunchai, B. Crémilleux, F. Jurie, Image re-ranking based on statistics of frequent patterns, in: Proceedings of the International Conference on Multimedia Retrieval, ACM, 2014, p. 129.
[40] J. Wan, D. Wang, S.C.H. Hoi, P. Wu, J. Zhu, Y. Zhang, J. Li, Deep learning for content-based image retrieval: a comprehensive study, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2014, pp. 157–166.
[41] G. Wang, D. Forsyth, Object image retrieval by exploiting online knowledge resources, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[42] J.Y. Wang, I. Almasri, X. Gao, Adaptive graph regularized nonnegative matrix factorization via feature selection, in: Proceedings of the International Conference on Pattern Recognition, IEEE, 2012, pp. 963–966.
[43] L. Wang, Z. Zhao, F. Su, Efficient multi-modal hypergraph learning for social image classification with complex label correlations, Neurocomputing 171 (2016) 242–251.
[44] M. Wang, W. Fu, S. Hao, D. Tao, X. Wu, Scalable semi-supervised learning by efficient anchor graph regularization, in: IEEE Transactions on Knowledge and Data Engineering, 28(7), IEEE, 2016, pp. 1864–1877.
[45] M. Wang, H. Li, D. Tao, K. Lu, X. Wu, Multimodal graph-based reranking for web image search, IEEE Trans. Image Process. 21 (11) (2012) 4649–4661.
[46] J. Weston, S. Bengio, N. Usunier, Large scale image annotation: learning to rank with joint word-image embeddings, Mach. Learn. 81 (1) (2010) 21–35.
[47] L. Wu, Y. Wang, J. Shepherd, Efficient image and tag co-ranking: a Bregman divergence optimization method, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2013, pp. 593–596.
[48] B. Xu, J. Bu, C. Chen, C. Wang, D. Cai, X. He, EMR: a scalable graph-based ranking model for content-based image retrieval, IEEE Trans. Knowl. Data Eng. 27 (1) (2015) 102–114.
[49] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, Y. Pan, A multimedia retrieval framework based on semi-supervised ranking and relevance feedback, IEEE Trans. Pattern Anal. Mach. Intell. 34 (4) (2012) 723–742.
[50] Y. Yang, D. Xu, F. Nie, J. Luo, Y. Zhuang, Ranking with local regression and global alignment for cross media retrieval, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2009, pp. 175–184.
[51] H. Zhang, Z.J. Zha, Y. Yang, S. Yan, Y. Gao, T.S. Chua, Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2013, pp. 33–42.
[52] S. Zhao, H. Yao, Y. Yang, Y. Zhang, Affective image retrieval via multi-graph learning, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2014, pp. 1025–1028.
[53] L. Zhu, J. Shen, H. Jin, R. Zheng, L. Xie, Content-based visual landmark search via multimodal hypergraph learning, IEEE Trans. Cybern. 45 (12) (2015) 2756–2769.
[54] X. Zhu, Q. Xie, Y. Zhu, X. Liu, S. Zhang, Multi-view multi-sparsity kernel reconstruction for multi-class image classification, Neurocomputing 169 (2015) 43–49.

Shuhan Qi received his M.S. degree in Computer Sciences from the Harbin Institute of Technology in 2011. Since 2011, he has been a Ph.D. degree candidate in Computer Sciences from Harbin Institute of Technology Shenzhen Graduate School. His research interests include computer vision, multimedia and pattern recognition.

Xuan Wang received his M.S. and Ph.D. degrees in Computer Sciences from Harbin Institute of Technology in 1994 and 1997, respectively. He is a professor and Ph.D. supervisor in the Computer Application Research Center, Harbin Institute of Technology Shenzhen Graduate School. His main research interests include artificial intelligence, computer vision, computer network security and computational linguistics.

Xi Zhang received his B.S. degree in School of Science from the Hubei University for Nationalities in 2013. Since 2014, he has been a M.S. degree candidate in Computer Sciences from Harbin Institute of Technology Shenzhen Graduate School. His research interests include computer vision, pattern recognition and mathematical modeling.

Xuemeng Song is currently a Ph.D. student with the school of computing, National University of Singapore. She received her degree from University of Science and Technology of China in 2012. Her research interests are information retrieval and social network analysis. She has published several papers in the top conferences and journals, such as SIGIR and TOIS.

Zoe L. Jiang received the Ph.D. degree from The University of Hong Kong, Hong Kong, in 2010. She is currently an Assistant Researcher with School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China. Her research interests include computer vision and pattern recognition.