International Journal of Mathematics Trends and Technology
ISSN: 2231 - 5373 /doi:10.14445/22315373/IJMTT-V67I3P508
Volume 67 Issue 3, 57-69, March, 2021
2021 Seventh Sense Research Group®

Collective Bayesian Matrix factorization Hashing for cross-modal retrieval

Loubna Karbil[1] and Ahmad Sani[2] and Imane Daoudi[3]

[1]National high school of Electricity and Mechanics, Hassan II University, Casablanca, Morocco

2Departement of Mathematics, Ibn Zohr University, Agadir, Morocco

3National high school of Electricity and Mechanics, Hassan II University, Casablanca, Morocco

ABSTRACT

Matrix factorization hashing approaches have been widely applied in large scale cross-modality visual search due to their efficiency to preserve similarities among multimodal features. In this paper, we propose a novel cross-modality hashing technic based on Bayesian matrix factorization that factorizes all modalities into a shared latent semantic space using the Bayesian inference. To achieve better search performance, we measure the similarity using the cosine distance. Several experiments prove that the proposed method achieves better performance than many known methods on cross-modal retrieval applications.

KEYWORDS

Cross-modal retrieval, Matrix Factorization, Bayesian hashing, hash function, Multimodal hashing

I.INTRODUCTION

During recent years, Cross-modal retrieval becomes much attractive mainly with the unprecedented blow up of multimedia data on the internet and social networks. Image-text search has a big role in many fields such as object recognition, video surveillance and audio-text recognition. The task of cross modal retrieval is to submit (1) (2). Cross modal retrieval can return more comprehensive results with information from different modalities compared to unimodal retrieval methods (3) (4).The key concept of multimodal hashing is transforming high multimodal data into a set of a small length compact code that preserves the cross-modality similarities of the original data . So, the multimodal similarity distance between the original data can be computed using a convenient metric between obtained codes. Faced to limited storage resource and the semantic gap between the representation and semantic label, makes performing accurate cross-modal retrieval on large scale datasets very challenging. Many approaches have been proposed to address the cross-modality retrieval problem and can be grouped into two categories:

Unsupervised methods that can learn the cross-modal feature representation space by maximizing the correlation between pairwise data from different modalities such as Co-Regularized Hashing (CRH) (8) that preserves the inter-modality similarity. In fact, the objective function consists in projecting simultaneously data from 0 for good generalization. Cross-view hashing (CVH) proposed by Kumar and Udupa (5), which extends the spectral hashing (SH) (6), the method reduces the dissimilarity between hash codes to preserve the distance between similar multimodal data modal. Zhu et al. (7) in linear cross-modal hashing (LCMH) used anchor maps to create similar hash code between similar data point intra modality, without constructing similarity maps

The advantage of Inter-media hashing (IMH) (9) is the fact it considers the differences between multiple modalities. It starts by exploring the correlations within each single modality according to similarity graph. Then it makes the binary codes of the multimodal data points relevant. Multimodal latent binary embedding (MLBE) (10) uses latent variable models to learn the binary hash codes. Multi-View Spectral Hashing (MVSH) (11) creates binary hash codes, using multiview information. Composite Hashing with Multiple Information Sources

(CHMIS) (12) integrates information from several different sources into the binary hashing codes by adjusting the weights on each individual source for maximizing the coding performance, and enables fast conversion from query examples to their binary hashing codes. (13) Propose a novel Latent Semantic Sparse Hashing (LSSH) to perform cross-modal similarity search by employing sparse coding to capture the salient structures of images and Matrix Factorization to learn the latent concepts from text. Semantic Topic Multimodal Hashing (STMH) (14) is developed by considering latent semantic information in coding procedure. It first discovers clustering patterns of texts and robust factorizes the matrix of images to obtain multiple semantic topics of texts and concepts of images. In the proposed RFDH model (15), binary codes are directly learned based on discrete matrix decomposition, so that the large quantization error caused by relaxation is avoided.

Supervised multimodal hashing methods can create hash codes with preserving semantic similarities between modalities by approximating the pairwise similarity matrices based on supervised information such as semantic correlation maximization (SCM) (16) which proposes to seamlessly integrate semantic labels into the hashing learning procedure for large-scale data modelling. Semantics-Preserving Hashing (SePH) method (17) proposed to transform semantic affinities of training data as supervised information into a probability distribution learned hash codes in Hamming space via minimizing the Kullback-Leibler divergence. Multimodal discriminative binary embedding (MDBE) (10) focused on learning discriminative hash codes by formulating the hash function learning in terms of classification, where the binary codes generated is discriminative and exploiting, the label information to discover the shared structures inside heterogeneous data. Cross Modal Hashing (CMDH) (18) learned a set of shared binary codes from different modalities, to remove the modality effectively in cross-modal multimedia retrieval. Supervised robust discrete multimodal hashing (SRDMH) (19) incorporates full label information into the hash functions learning to preserve the similarity in the original space. Supervised hierarchical cross-modal hashing (HiCHNet) (20) exploits the hierarchical labels of instances by the pre-established label hierarchy and characterize each modality of the instance with a set of layer-wise hash representations. Recently, favourable performance can be achieved using deep supervised multimodal hashing methods thanks to using deep neural networks while constructing the hash code. Those algorithms can highly capture nonlinear correlations between multimodal data such as deep semantic multimodal hashing network (DSMHN) (21), Cross-modal hashing with semantic deep embedding (CMSDE) (22) and Pairwise Correlation Discrete Hashing (PCDH) (23). We can conclude that supervised methods needs usually supervised information of the training data set which makes performing accurate cross-modal retrieval on large scale datasets very challenging.

In the current work, we develop a new Bayesian specific modality Matrix factorization hashing method which improves matrix factorization hashing using a fully Bayesian treatment of the probabilistic matrix factorization. We summarize the main contributions of our algorithm as follows:

1. The use of probabilistic matrix factorization learning algorithm improves the initialization of projection matrix in steed of using random ones.

2. The representation of the training data is adapted by subtracting the error found between different unified representations of modalities.

3. To have higher performance, we use the cosine distance to measure similarity.

4. The fully Bayesian inference model used in our methods significantly increases the model's predictive accuracy.

5. Experimental results on two real-world datasets show that our method significantly outperforms the state of the art multimodal methods in terms of both accuracy and scalability.

II. Related work

In (24), Collective Matrix Factorization Hashing (CMFH) learns unified hash codes by collective matrix factorization with latent factor model from different modalities as one instance. The CMFH+ in (25) improves the collaborative matrix factorization by adapting the representation of the training data with subtracting the

error found between different unified representations of modalities.(26) proposes supervised matrix factorization hashing with quantitative loss (SMFH-QL) which generates hash codes via the class label and use matrix factorization to design hash codes from original multimodal data. Modality-specific Matrix Factorization Hashing (MsMFH) (27) factorizes the original feature representations into individual latent semantic representations, and then align those representations using an orthogonal transformation. Joint and individual matrix factorization hashing (JIMFH) (28) method learns unified hash codes using collective matrix factorization, and uses individual matrix factorization to get individual hash codes.

In (29) they presented the Probabilistic Matrix Factorization (PMF) model which scales linearly with the number of observations and, more importantly, performs well on the large, sparse, and very imbalanced Netflix dataset. To improve collaborative filtering in Netflix dataset, (30) presented a Bayesian treatment of the Probabilistic Matrix Factorization (PMF) model in which model capacity is controlled automatically by integrating over all model parameters and hyper parameters trained using Markov chain Monte Carlo methods. (31) proposes an adaptive initialisation to the Bayesian matrix factorization which still suffers from the cold-start problem where predictions of ratings for new items or of new user's preferences are required in Netflix dataset.

Inspired by those works, this paper focuses on leaning modality specific latent representation using Bayesian inference to solve collective matrix factorization between different modalities and learn hash function which is to the best our knowledge not done before.

III. Proposed method

In this work and for ease of presentation, here we describe the method problem with only two modalities (image modality and text modality), which can be easily extended to cases with more than two modalities.

III.1. Notation and problem formulation

Table 1 shows notations and their signification respectively used to study this paper. Our algorithm accepts paired image-text data for input and processed to a specific treatment using hash coding to retrieve at the at the end the most relevant texts given an image query and vice versa.

Table 1. Notation used in this paper.

Notation	Definition
O	set of multimodal objets
X_1, X_2	Image Modality, Text Modality
U_1,U_2	Basic matrices for matrix factorization
W_1,W_2	Hash functions for matrix factorization
V	Shared latent representation
N	number of training objects)
d_1,d_2	Dimension of image, text modality dataset
k	Number of latent factors

III.2. Framework overview

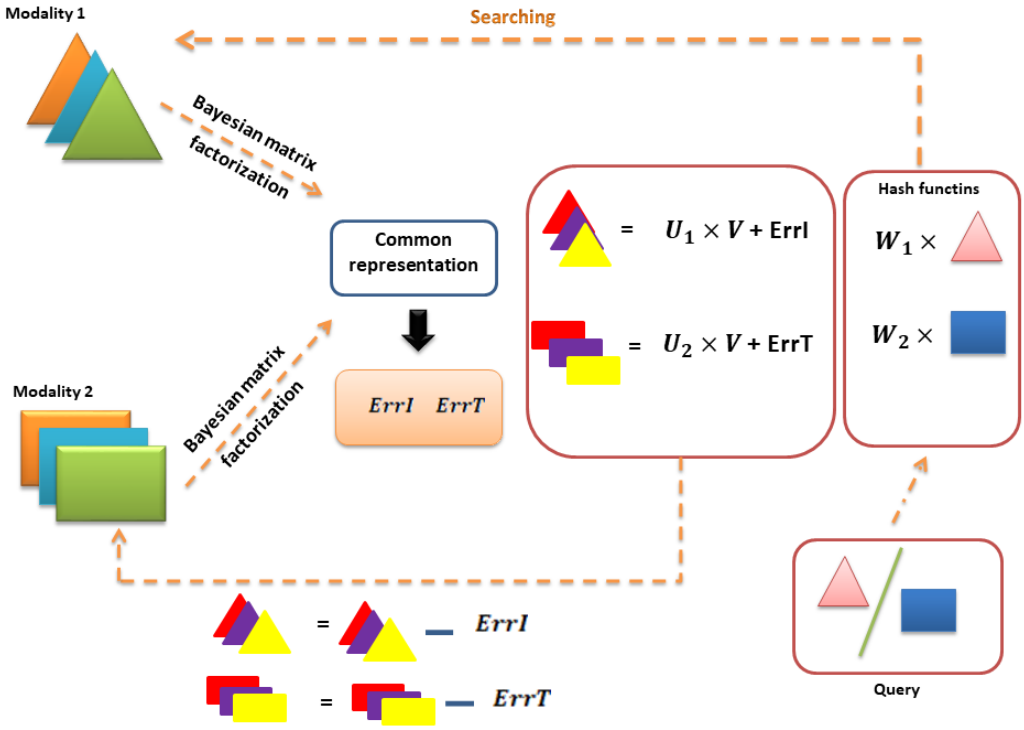

Figure 1. Framework of Our method, illustrated with toy data

We learn semantic features from data in the Collective Matrix Factorization Hashing (24) by using matrix factorization:

$$X^{(t)} = U_t V_t \,, \quad \forall t \in \{1,2\} \qquad (1)$$

Where $U_t \in \mathbb{R}^{d_t \times k}, V_t \in \mathbb{R}^{k \times n}$. Each column vector v_t is a latent semantic representation of the t-th view data $x^{(t)}$.

Assuming that the interconnected data should have the identic latent semantic representation, $X^{(1)}$ and $X^{(2)}$ are factorized together with the constraint $V_1 = V_2 = V$, where v_i is a column vector that can be used as a semantic hash code for the object data o_i , $i = \{1,2,\dots,n\}$ from database.

When we have a new input, we extend the method by learning hash functions like (1).

$$V = W_t X^t \; \forall \; t \in \{1,2\} \qquad (2)$$

We combine in the overall objective function data matrix factorization (a), projection matrix (b) and regularization term (c):

$$\begin{aligned} G = {} & \lambda \|X^1 - U_1 V\|_F^2 + (1-\lambda)\|X^2 - U_2 V\|_F^2 && \text{(a)} \\ & + \delta(\|V - W_1 X^1\|_F^2 + \|V - W_2 X^2\|_F^2) && \text{(b)} \\ & + \gamma(\|V\|_F^2 + \|U_1\|_F^2 + \|U_2\|_F^2 + \|W_1\|_F^2 + \|W_2\|_F^2) && \text{(c)} \end{aligned} \qquad (3)$$

Where λ is the balance parameter, δ and γ are trade-off parameters. Our method have the objective of minimizing G and solve $U_1,\ U_2,\ V\ ,\ W_1$ and W_2 using the Bayesian inference.

III.3. Learning Hash Function

III.3.1. Bayesian inference

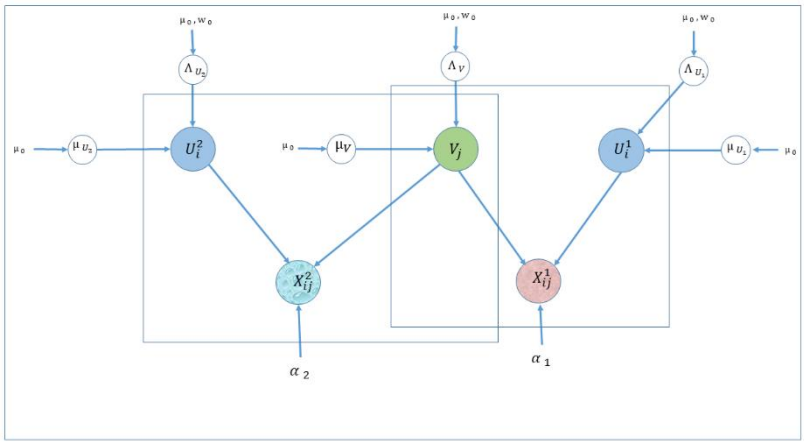

Figure 2. Graphical model of the Bayesian matrix factorization

The graphical model representing the collective Bayesian matrix factorization is shown in Fig. 1. The prior distributions over U_t and V_t are assumed to be Gaussian. So the likelihood will be given like (30) by

$$P(U_t|\mu_{U_t}, \Lambda_{U_t}) = \prod_{i=1}^{N} \mathcal{N}\left(U_t^i|\mu_{U_t}, {\Lambda_{U_t}}^{-1}\right) \quad (4)$$

$$P(V_t|\mu_{V_t}, \Lambda_{V_t}) = \prod_{i=1}^{N} \mathcal{N}\left(V_t^i|\mu_{V_t}, {\Lambda_{V_t}}^{-1}\right) \quad (5)$$

With the constraint $V_1 = V_2$

We define the Gaussian-Wishart priors on the text and image hyperparameters

$\theta_{U_t} = \{\mu_{U_t}, \Lambda_{U_t}\}$ and $\theta_{V_t} = \{\mu_{UV_t}, \Lambda_{V_t}\}$:

$$P(\theta_{U_t}|\theta_0) = p(\mu_{U_t}|\Lambda_{U_t})p(\Lambda_{U_t}) = \mathcal{N}\left(\mu_{U_t}\middle|\mu_0, (\beta_0\Lambda_{U_t})^{-1}\right)\mathcal{W}(\Lambda_{U_t}|W_0, \nu_0) \quad (6)$$

$$P(\theta_{V_t}|\theta_0) = p(\mu_{V_t}|\Lambda_{V_t})p(\Lambda_{V_t}) = \mathcal{N}\left(\mu_{V_t}\middle|\mu_0, (\beta_0\Lambda_{V_t})^{-1}\right)\mathcal{W}(\Lambda_{V_t}|W_0, \nu_0) \quad (7)$$

Where $\mathcal{W}$ is the Wishart distribution with ν_0 degrees of freedom and:

$$\mathcal{W}(\Lambda|W_0, \nu_0) \propto |\Lambda|^{\frac{(\nu_0 - D - 1)}{2}} \exp\left(-\frac{1}{2}Tr\left({W_0}^{-1}\Lambda\right)\right)$$

For convenience we define $\theta_0 = \{\mu_0, \nu_0, \mathcal{W}_0\}$

For ease of addressing the problem, we will use the Gibbs sampling (32) algorithm to resolve the equation (4) and (5).

Gibbs sampling (32) consists of looping through the latent variables, sampling each one from its distribution conditional on the current values of all other variables.

We used a conjugate priors for the parameters and hyperparameters in our collective Bayesian matrix factorisation model which made sampling the conditional distribution over different modalities becomes easy

$$P\left(U_t^i\middle|X^t, V, \theta_{U_t}, \alpha\right) = \mathcal{N}\left(U_t^i\middle|\mu_t^{i*}, \left[\Lambda_t^{i*}\right]^{-1}\right) \quad (8)$$

$$\sim \prod_{j=1}^{M}\left[\mathcal{N}\left(X_{ij}^t\middle|{U_t^i}^T V^j, \alpha^{-1}\right)\right]^{I_{ij}} P(U_t|\mu_{U_t}, \Lambda_{U_t})$$

Where

$$\Lambda_t^{i*} = \Lambda_{U_t} + \alpha \sum_{j=1}^{M} [V_j V_j^T]^{I_{ij}} \quad (9)$$

$$\mu_t^{i*} = [\Lambda_t^{i*}]^{-1} (\alpha \sum_{j=1}^{M} [V_j X_{ij}^t]^{I_{ij}} + \Lambda_{U_t} \mu_{U_t} \quad (10)$$

Notice that the conditional distribution over the latent feature matrix U_t (idem for V_t) factorizes into the product of conditional distributions over the individual feature vectors

$$p(U_t | X^t, V, \theta_{U_t}) = \prod_{i=1}^{N} p(U_t^i | X^t, V, \theta_{U_t})$$

Which allows us to sample those conditional distribution in parallel and speed up the sampler, this make the algorithm less time consuming in large scale datasets.

The conditional distribution over modalities hyperparameters conditioned on features feature matrix U_t (idem for V_t) is given by the Wishart-Gaussian distribution:

$$p(\mu_{U_t}, \Lambda_{U_t} | U_t, \theta_0) = \mathcal{N}\left(\mu_{U_t} | \mu_0^*, (\beta_0^* \Lambda_{U_t})^{-1}\right) \mathcal{W}(\Lambda_{U_t} | \mathcal{W}_0^*, \nu_0^*) \quad (11)$$

Where:

$$\mu_0^* = \frac{\beta_0 \mu_0 + NH}{\beta_0 + N}, \qquad \beta_0^* = \beta_0 + N, \qquad \nu_0^* = \nu_0 + N,$$

$$[\mathcal{W}_0^*]^{-1} = W_0^{-1} + NL + \frac{\beta_0 N}{\beta_0 + N}(\mu_0 - H)(\mu_0 - H)^T,$$

$$H = \frac{1}{N}\sum_{i=1}^{N} U_t^i, \qquad L = \frac{1}{N}\sum_{i=1}^{N} U_t^i U_t^{iT},$$

With the additional condition that $V_1 = V_2 = V$.

III.3.2 The proposed algorithm

The solution of the optimization problem is done by sampling the Bayesian inference over the modalities matrix using the Gibbs sampling algorithm like in (30) . We minimize the objective function using eight steps.

Our method relies on eight steps that are :

1- Initialize $\mathbf{U_t}^1, \mathbf{V_t}^1$ using the probabilistic matrix factorization in (29).

2- Sample the hyperparameters k times

$$\theta_{U_t}^k \sim p(\theta_{U_t} | {U_t}^k, \theta_0) \quad \forall \mathbf{t} \in \{\mathbf{1}, \mathbf{2}\} \quad (12)$$

$$\theta_{V_t}^k \sim p(\theta_{V_t} | {V_t}^k, \theta_0) \quad \forall \mathbf{t} \in \{\mathbf{1}, \mathbf{2}\} \quad (13)$$

3- For every sampling of the hyperparameters we calculate a new iteration of U_t matrix in parallel

$$U_t^{i\,k+1} \sim P(U_t^i | X_t, {V_t}^k, {\theta_{U_t}}^k) \quad \forall \mathbf{t} \in \{\mathbf{1}, \mathbf{2}\} \quad (14)$$

4- For every sampling of the hyperparameters we calculate a new iteration of V_t matrix in parallel

$$V_t^{i\,k+1} \sim P(U_t^i | X^t, {U_t}^{k+1}, {\theta_{V_t}}^k) \quad \forall \mathbf{t} \in \{\mathbf{1}, \mathbf{2}\} \quad (15)$$

5- Calculate U_t and V_t until $|V_1 - V_2| < \varepsilon \quad \forall \mathbf{t} \in \{\mathbf{1}, \mathbf{2}\}$

6- Measure the projection error matrix like :

$$\mathbf{Err^t = X^t - U_t V} \quad \forall \mathbf{t} \in \{1, 2\} \qquad (16)$$

7- Data representation of the training data is adjusted when we substruct the equation (16) from $\mathbf{X^t}$ like:

$$\mathbf{X^t = X^t - Err^t} \quad \forall \mathbf{t} \in \{1, 2\} \qquad (17)$$

8- Fixing V, we calculate W_1 and W_2 using ascent method for linear triangular systems.

The algorithm is summarized in figure 2

Algorithm 1 Initialization and Indexing

Input:

Data matrix X^t where t=1, 2.

Output:

Hash codes $\boldsymbol{V}$.
Projection matrix $\boldsymbol{W_t}$
Data representation $\boldsymbol{X^t}$

Repeat

1- Initialize model parameters $\mathbf{U_t}^1, \mathbf{V_t}^1$ where t=1, 2

2- For k =1 to M

- Sample the hyperparameters

$$\theta_{U_t}^{k} \sim p\left(\theta_{U_t} | {U_t}^k, \theta_0\right) \quad \forall \mathbf{t} \in \{1, 2\}$$

$$\theta_{V_t}^{k} \sim p\left(\theta_{V_t} | {V_t}^k, \theta_0\right) \quad \forall \mathbf{t} \in \{1, 2\}$$

- For i=1 to $\boldsymbol{d_t}$ where t=1, 2

  Calculate a new iteration of U_t matrix in parallel using equation 14

- For i=1 to N

  Calculate a new iteration of V_t matrix in parallel using equation 15

While abs $|\boldsymbol{V_1 - V_2}| < \boldsymbol{\varepsilon}$

3- Calculate W_t , t =1, 2.

4- Calculate Err^t using equation (16), t =1, 2.

5- Calculate X^t using equation (17), t =1, 2.

6- Return W_t , V , X^t.

Figure 3. Initialization and indexing algorithm.

III.3.3 Searching algorithm

Our method uses the Bayesian inference to compute the latent semantic vectors V. We consider each V_j as a hash code of the j-th data from the database.

We generate the hash code for a new query like:

$$v_q = W_t \times (q^t - \sum \frac{x_i^t}{n}) \qquad (18)$$

We calculate the similarity between two vectors X and Y using the cosine distance like:

$$D_C(X,Y) = 1 - \frac{XY}{\|X\|\|Y\|} \qquad (19)$$

The searching algorithm is summarized in figure 3.

Algorithm 2 Searching

Input

W_t ,V , q^t and X^t t=1, 2.

Output

K-NN data points

1- Train q^t projection by equation (18), t=1, 2.

2- Measure the Cosine similarity between V and v_q using equation (19).

3- Retrieve the K-NN from the new X^t.

Figure 4. Searching algorithm.

IV.Experiments

We realised experiments on two real world multimodal datasets, we are used to check the efficiency of our hashing method. All our experiments are carried out on a computer with Intel (R) CPU X7560@2.27 GHz and 22 GB RAM.

IV.1. Datasets

NUS–WIDE (33): is a multiclass dataset containing 269,648 annotated images from Flickr. Images are manually classified into 81 classes (an image can belong to more than one class) and represented as bags of 500 dimensions of SIFT (BoF, used as the first modality) and bags of 1000 dimensions of text tags. (Tags, used as second modality). We select 186,577 image-tag pairs that belong to the 10 largest concepts.

Wiki (34): is explored from featured Wikipedia articles. It consists of 2,866 documents which are image-text pairs and annotated with semantic tags of 10 classes. a 128-dimensional SIFT feature vector represents each image and each text is represented by a 10-dimensional topic vector.

IV.2. Experiments results

IV.2.1 Evaluation metrics

We evaluate the retrieval performance using the mean average precision (MAP) which is defined like:

$$MAP = \frac{1}{L}\sum_{r=1}^{R} P(r) \times rel(r) \qquad (12)$$

Where $rel(r) = \begin{cases} 1 , & if\ the\ rank\ is\ relevent \\ 0 & ,otherwise \end{cases}$

And R, P(r) and L denote respectively the number of retrieved results, the percentage of relevant results in the first r top-ranked retrieved matches and the number of ground-truth neighbours of a query.

We set $v_0 = k$ and $\mathcal{W}_0$ $k \times k$ identity matrix for both U_t and V_t then we choose for the convenience $\mu_0 = 0$ by symmetry. We loop the Gibbs sampling 2 times per experiment, we set the observation noise $\alpha_1 = \alpha_2 = \alpha = 2$.

IV .2.2 Rival Methods

Our method is compared on two cross modal retrieval tasks: querying, mutatis mutandis, image database by text sets. We evaluate our method with several known and recent multimodal hashing methods: CVH (5) , SMH (35) , CCA (36) , SMFH , CMFH , CMFH+ , LSSH , SMFH (26) , SCM (16), DCH((23), LCMFH (7), EDSH (37) , MSMFH (27) . We carefully adjusted their parameters and reported the best of their results.

IV.4. Experiments results

IV.4.1 Results on Wiki

As usual, for the Wiki dataset 80% of the data is used as training set and 20% as the query set. We reported the MAP values for our method and for all rival methods on Table 2 under setting of 16, 32, 64 and 128 bits respectively. The Table 2 evaluates the proposed algorithm on wiki data set. This evaluation significantly shows that the CBMFH achieved better performance on this benchmark of both text to image and image to text sides, and the performance gain of the proposed method is more than 10%. CBMFH has done remarkable MAP. Compared to the rival methods on both views and where varying the code length, our method is more efficient especially in the text to image task which is due to the fact that the Bayesian matrix factorization can successfully find better latent semantic features from text.

Table 3. MAP results on WIKI dataset.

Task (request)	Method	Wiki			
		16 bits	32 bits	64 bits	128 bits
Image To Text	CVH	0,2041	0,1604	0,1296	0,1308
	SMH	0,2257	0,2459	0,249	0,2555
	CCA	0,1578	0,1393	0,1282	0,1237
	CMFH	0,2437	0,2523	0,2575	0,2611
	CMFH+	0,2642	0,2675	0,2675	0,2675
	LSSH	0,2297	0,2532	0,2406	0,2336
	SMFH	0,2276	0,2514	0,2533	0,2564
	SCM	0,2474	0,2514	0,2533	0,2564
	DCH	0,3423	0,3599	0,3806	0,3848
	LCMFH	0,3114	0,3337	0,3455	0,3606
	EDSH	0,3871	0,3942	0,3950	0,3956
	MSMFH	0,3832	0,3953	0,4045	0,4050
	Our	**0,4133**	**0,4135**	**0,4136**	**0,4136**
Text To Image	CVH	0,2962	0,1944	0,1337	0,1125
	SMH	0,2341	0,241	0,2458	0,26
	CCA	0,1755	0,1619	0,155	0,155
	CMFH	0,6107	0,6282	0,6376	0,6471
	CMFH+	0,6588	0,6558	0,6558	0,6558
	LSSH	0,6134	0,6296	0,6349	0,6304
	SMFH	0,5590	0,6473	0,6678	0,6617
	SCM	0,3819	0,4479	0,4418	0,4405
	DCH	0,6999	0,7051	0,7117	0,7222
	LCMFH	0,6969	0,7042	0,7094	0,7408
	EDSH	0,5581	0,5661	0,5662	0,5732
	MSMFH	0,5891	0,6082	0,6183	0,6221
	Our	**0,7525**	**0,7528**	**0,7531**	**0,7532**

From results above, it is easy to see that our method has an almost the same MAP result over the variation of the code length which gives stability to our algorithm. When we retrieve the factorization error from data representation and we use the new representation to search relevant data which makes the performance of our algorithm more stable under the variation of the code length.

IV.4.2 Results on NUS-WIDE

Despite the cost in terms of training time of rival methods on NUS-WIDE, we evaluated and compared our algorithm to several recent methods in the literature. Better MAP result gains are obtained on the large scale NUS-WIDE, mainly in the Text to image retrieval. Table 3 shows that our method is more efficient in terms of search accuracy and precision. In addition, we can observe that our method remains invariant under the variation of the code length. This is an obvious consequence since we reinject the factorization error and we change the initial data representation. We can improve the research performance of the method by using cosine distance.

Table 4. MAP results on NUS-WIDE dataset.

Task (request)	Method	NUS-WIDE			
		16 bits	32 bits	64 bits	128 bits
Image To Text	CVH	0,5216	0,5146	0,4912	0,4600
	SMH	0.2918	0.3004	0.3235	0.3225
	CCA	0.3561	0.3451	0.3411	0.3439
	CMFH	0.3950	0.3859	0.3741	0.3891
	CMFH+	0.4054	0.4054	0.4054	0.4054
	LSSH	0,5047	0,5176	0,5229	0,5439
	SMFH	0,5671	0,6031	0,6117	0,6282
	SCM	0,5496	0,5642	0,5425	0,5362
	DCH	0,5686	0,6189	0,6118	0,6272
	LCMFH	0,6191	0,6166	0,6203	0,6295
	EDSH	0,5141	0,5211	0,5892	0,6073
	MSMFH	0,5521	0,5641	0,5894	0,6072
	Our	0,6561	0,6562	0,6572	0,6572
Text To Image	CVH	0,5462	0,5311	0,4968	0,4617
	SMH	0.3552	0.3459	0.3571	0.3619
	CCA	0.3550	0.3516	0.3489	0.3476
	CMFH	0.4175	0.4046	0.3929	0.4048
	CMFH+	0.5163	0.5163	0.5163	0.5163
	LSSH	0,6201	0,6494	0,6841	0,7055
	SMFH	0,5843	0,6522	0,6750	0,7163
	SCM	0,5496	0,5691	0,5755	0,5955
	DCH	0,7376	0,7815	0,7898	0,8036
	LCMFH	0,7052	0,6961	0,7074	0,7178
	EDSH	0,6341	0,6561	0,6693	0,6834
	MSMFH	0,6792	0,6954	0,7121	0,7163
	Our	0,7934	0,7956	0,7958	0,7958

For data with small code length, our method seems more efficient than the state of the art method. Despite that DCH achieves better performance on 128 bits code length, in general our method remains more accurate especially when the latent semantic representation is small enough.

IV.4.3 Scalability

This subsection is devoted to evaluate the scalability of different methods compared to ours. We start by experimenting the training time of different baseline methods on NUS-WIDE dataset when we vary the code length. Results are summarized in table 4. The benchmark is done over the methods of the state of the art that have good MAP performance: LSSH, DCH, CMFH, CMFH+ and LCMFH.

Table 4. Training time (in second) on NUS-WIDE dataset under the code length variation.

Method	NUS-WIDE			
	16 bits	32 bits	64 bits	128 bits
LSSH	762.61	895.42	877.88	889.08
DCH	12.82	12.63	14.36	17.88
CMFH	545.95	591.12	522.85	703.60
CMFH+	550.01	592.13	519.33	668.24
LCMFH	5.44	5.18	6.69	8.92
Ours	18.21	18.9	19.12	21.54

From table 4 we can see that LSSH, CMFH and CMFH+ require much training time than DCH, LCMFH and our method. Additionally, the last three have a good training times despite that LCMFH achieved the best results than DCH and our method. However in summary, the proposed method shows the best retrieval performance in a competitive time.

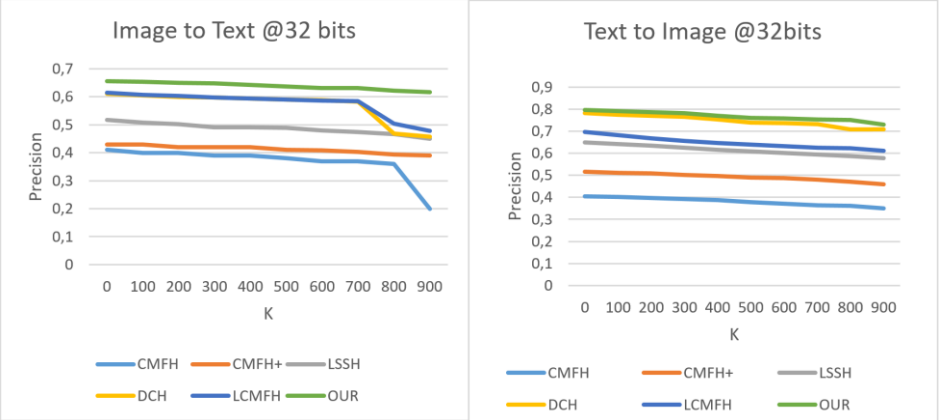

Figure 3: the top K-Precision curves for all the baseline methods

Figure 3 show the top K-Precision curves for all the baseline method respectively. The code length is fixed to 32 in this experiment. According to the experimental results, our proposed CBMFH outperforms the other methods in both image-text and text-image search tasks because of the efficiency of resolving the matrix factorization using the Bayesian inference and the retrieving from data representation the factorization error and search from the new representation of data.

V. Conclusion

We have proposed a new method relied on Collaborative Bayesian Matrix Factorization and have improved the precision by reducing the error between multimodal data and their hash code. The method exposed here is based on a probabilistic matrix factorization algorithm to initialize projection matrix. The similarity was measured using the cosine distance. Several experiments on two real datasets have proved that our method may significantly outperform and probably improve the state of the art methods in term of scalability and accuracy.

VI.RÉFÉRENCES

1- W. Wang, X. Yang, B. C. Ooi, D. Zhang, and Y. Zhuang, "Effectivedeep learning-based multi-modal retrieval," VLDBJ, vol. 25, no. 1, pp.79–101, 2016.

2- PEREIRA, Jose Costa, COVIELLO, Emanuele, DOYLE, Gabriel, et al. On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE transactions on pattern analysis and machine intelligence, 2013, vol. 36, no 3, p. 521-535.

3- W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang. Supervised hashing with kernels. In CVPR. IEEE, 2012.

4- B. Kulis and K. Grauman. Kernelized locality-sensitive hashing for scalable image search. In ICCV. IEEE, 2009.

5- KUMAR, Shaishav et UDUPA, Raghavendra. Learning hash functions for cross-view similarity search. In : Twenty-second international joint conference on artificial intelligence. 2011.

6- WEISS, Yair, TORRALBA, Antonio, FERGUS, Robert, et al. Spectral hashing. In : Nips. 2008. p. 4.

7- ZHU, Xiaofeng, HUANG, Zi, SHEN, Heng Tao, et al. Linear cross-modal hashing for efficient multimedia search. In : Proceedings of the 21st ACM international conference on Multimedia. 2013. p. 143-152.

8- ZHEN, Yi et YEUNG, Dit Yan. Co-regularized hashing for multimodal data. Advances in neural information processing systems, 2012, vol. 2, p. 1376.

9- SONG, Jingkuan, YANG, Yang, YANG, Yi, et al. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In : Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. 2013. p. 785-796.

10- WANG, Di, GAO, Xinbo, WANG, Xiumei, et al. Multimodal discriminative binary embedding for large-scale cross-modal retrieval. IEEE Transactions on Image Processing, 2016, vol. 25, no 10, p. 4540-4554.

11- SHEN, Xiaobo, SHEN, Fumin, SUN, Quan-Sen, et al. Multi-view latent hashing for efficient multimedia search. In : Proceedings of the 23rd ACM international conference on Multimedia. 2015. p. 831-834.

12- ZHANG, Dan, WANG, Fei, et SI, Luo. Composite hashing with multiple information sources. In : Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 2011. p. 225-234.

13- ZHOU, Jile, DING, Guiguang, et GUO, Yuchen. Latent semantic sparse hashing for cross-modal similarity search. In : Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. 2014. p. 415-424.

14- WANG, Di, GAO, Xinbo, WANG, Xiumei, et al. Semantic topic multimodal hashing for cross-media retrieval. In : Twenty-fourth international joint conference on artificial intelligence. 2015.

15- WANG, Di, WANG, Quan, et GAO, Xinbo. Robust and flexible discrete hashing for cross-modal similarity search. IEEE Transactions on Circuits and Systems for Video Technology, 2017, vol. 28, no 10, p. 2703-2715.

16- ZHANG, Dongqing et LI, Wu-Jun. Large-scale supervised multimodal hashing with semantic correlation maximization. In : Proceedings of the AAAI conference on artificial intelligence. 2014.

17- LIN, Zijia, DING, Guiguang, HU, Mingqing, et al. Semantics-preserving hashing for cross-view retrieval. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 3864-3872.

18- JIANG, Qing-Yuan et LI, Wu-Jun. Deep cross-modal hashing. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 3232-3240.

19- LI, Chuan-Xiang, YAN, Ting-Kun, LUO, Xin, et al. Supervised robust discrete multimodal hashing for cross-media retrieval. IEEE Transactions on Multimedia, 2019, vol. 21, no 11, p. 2863-2877.

20- SUN, Changchang, SONG, Xuemeng, FENG, Fuli, et al. Supervised hierarchical cross-modal hashing. In : Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019. p. 725-734.

21- JIN, Lu, LI, Zechao, et TANG, Jinhui. Deep semantic multimodal hashing network for scalable image-text and video-text retrievals. IEEE Transactions on Neural Networks and Learning Systems, 2020.

22- YAN, Cheng, BAI, Xiao, WANG, Shuai, et al. Cross-modal hashing with semantic deep embedding. Neurocomputing, 2019, vol. 337, p. 58-66.

23- CHEN, Yaxiong et LU, Xiaoqiang. Deep discrete hashing with pairwise correlation learning. Neurocomputing, 2020, vol. 385, p. 111-121.

24- DING, Guiguang, GUO, Yuchen, et ZHOU, Jile. Collective matrix factorization hashing for multimodal data. In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p. 2075-2082.

25- KARBIL, Loubna et DAOUDI, Imane. Large-Scale Supervised Hashing for Cross-Modal Retreival. In : 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA). IEEE, 2017. p. 803-808.

26- ZHAO, Huan, WANG, Song, SHE, Xiaolin, et al. Supervised Matrix Factorization Hashing With Quantitative Loss for Image-Text Search. IEEE Access, 2020, vol. 8, p. 102051-102064.

27- XIONG, Haixia, OU, Weihua, YAN, Zengxian, et al. Modality-specific matrix factorization hashing for cross-modal retrieval. Journal of Ambient Intelligence and Humanized Computing, 2020, p. 1-15.

28- WANG, Di, WANG, Quan, HE, Lihuo, et al. Joint and individual matrix factorization hashing for large-scale cross-modal retrieval. Pattern Recognition, 2020, vol. 107, p. 107479.

29- MNIH, Andriy et SALAKHUTDINOV, Russ R. Probabilistic matrix factorization.

30- Advances in neural information processing systems, 2007, vol. 20, p. 1257-1264.

31- SALAKHUTDINOV, Ruslan et MNIH, Andriy. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In : Proceedings of the 25th international conference on Machine learning. 2008. p. 880-887.

32- GANTNER, Zeno, DRUMOND, Lucas, FREUDENTHALER, Christoph, et al. Learning attribute-to-feature mappings for cold-start recommendations. In : 2010 IEEE International Conference on Data Mining. IEEE, 2010. p. 176-185.

33- GELFAND, Alan E. Gibbs sampling. Journal of the American statistical Association, 2000, vol. 95, no 452, p. 1300-1304.

34- CHUA, Tat-Seng, TANG, Jinhui, HONG, Richang, et al. Nus-wide: a real-world web image database from national university of singapore. In : Proceedings of the ACM international conference on image and video retrieval. 2009. p. 1-9.

http://www.svcl.ucsd.edu/projects/crossmodal/.

35- RASIWASIA, Nikhil, COSTA PEREIRA, Jose, COVIELLO, Emanuele, et al. A new approach to cross-modal multimedia retrieval. In : Proceedings of the 18th ACM international conference on Multimedia. 2010. p. 251-260.

36- HARDOON, David R., SZEDMAK, Sandor, et SHAWE-TAYLOR, John. Canonical correlation analysis: An overview with application to learning methods. Neural computation, 2004, vol. 16, no 12, p. 2639-2664.

37- YAO, Tao, HAN, Yaru, WANG, Ruxin, et al. Efficient discrete supervised hashing for large-scale cross-modal retrieval. Neurocomputing, 2020, vol. 385, p. 358-367.

38- WANG, Di, GAO, Xinbo, WANG, Xiumei, et al. Multimodal discriminative binary embedding for large-scale cross-modal retrieval. IEEE Transactions on Image Processing, 2016, vol. 25, no 10, p. 4540-4554.

39- ZHU, Xiaofeng, HUANG, Zi, SHEN, Heng Tao, et al. Linear cross-modal hashing for efficient multimedia search. In : Proceedings of the 21st ACM international conference on Multimedia. 2013. p. 143-152.