Cluster-based Joint Matrix Factorization Hashing for Cross-Modal Retrieval

Dimitrios Rafailidis
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
draf@csd.auth.gr

Fabio Crestani
Faculty of Informatics
Università della Svizzera italiana (USI)
Lugano, Switzerland
fabio.crestani@usi.ch

ABSTRACT

Cross-modal retrieval has been an emerging topic over the last years, as modern applications have to efficiently search for multimedia documents with different modalities. In this study, we propose a cross-modal hashing method by following a cluster-based joint matrix factorization strategy. Our method first builds clusters for each modality separately and then generates a cross-modal cluster representation for each document. We formulate a joint matrix factorization process with the constraint that pushes the documents' representations of the different modalities and the cross-modal cluster representations into a common consensus matrix. In doing so, we capture the inter-modality, intra-modality and cluster-based similarities in a unified latent space. Finally, we present an efficient way to generate the hash codes using the maximum entropy principle and compute the binary codes for external queries. In our experiments with two publicly available data sets, we show that the proposed method outperforms state-of-the-art hashing methods for different cross-modal retrieval tasks.

CCS Concepts

•Information systems → Multimedia information systems;

Keywords

Hashing; cross-modal retrieval; matrix factorization

1. INTRODUCTION

In multimedia applications, hashing techniques have been widely used for large-scale similarity search, such as locality sensitive hashing [4], iterative quantization [5] and spectral hashing [8]. The key idea is to design hash functions and learn similarity preserving binary codes for data representation with low storage cost and fast query speed. In the aforementioned hashing methods, given a query from one modality, for example an image-query, results are efficiently retrieved from an image database. In this study, we focus on cross-modal hashing, for example, given a query from an image modality, how to return the most relevant results from a textual modality. The rich representation of a document with different modalities has many applications; for instance, in a real-world multimedia application, given an image-query, a search engine can return relevant text documents to describe its details.

To handle the large amount of available multimedia content with different modalities in modern applications, several cross-modal hashing methods have been proposed [1, 6, 11, 13]. The main challenge of cross-modal hashing is how to learn the binary codes by capturing both the inter-modality and intra-modality similarities [2, 12]. For instance, [1] constructs a unified space and learns groups of hash functions with eigendecomposition and AdaBoost to ensure that if two documents with different modalities are relevant, then their corresponding hash codes are similar. [6] extends spectral hashing to cross-modal retrieval by formulating the learning of binary codes as a tractable eigenvalue problem. [11] presents an iterative scheme to learn a shared hamming space using a graph regularization formulation and a set of binary classifiers. [13] constructs heterogeneous hamming spaces and then connect them, while preserving the local structure using an anchor-based representation. Finally, [12] considers inter-modality and intra-modality consistency to generate a common hamming space, and integrates a linear regression model to learn hash functions so that the binary codes for new documents can be efficiently generated.

In [2], authors make the first attempt to compute hash codes using joint/collective matrix factorization with a latent factor model. Each matrix corresponds to the documents' representations for each modality, and then by jointly factorizing the matrices unified hash codes are generated. By revealing the associations between the different modalities and computing the similarities in each modality, the joint matrix factorization hashing method of [2] achieves high cross-modal retrieval accuracy. Nonetheless, the baseline joint matrix factorization that is used in [2] does not preserve the similarities that documents have in their neighborhoods/clusters in the same modality, nor considers the associations that the neighborhoods have in the different modalities.

To capture the inter-modality and intra-modality similarities at a neighborhood-based level, in this study we introduce a cluster-based joint matrix factorization method for cross-modal hashing. Our method consists of three steps, firstly we propose an efficient way to generate cross-modal

SIGIR '16, July 17-21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2914710

cluster representations for the documents. In the second step, we incorporate the generated cross-modal cluster representations into a joint factorization technique to compute the inter-modality and intra-modality similarities of the documents in the different modalities. This is achieved by formulating a joint matrix factorization process with the constraint that pushes the documents' representations of the different modalities and the cross-modal cluster representations into a common consensus matrix. Finally, we calculate the unified hash codes based on the maximum entropy principle [8], as well as the projection matrices to generate hash codes for external documents that do not belong to the database. Our experiments on two benchmark data sets demonstrate the superiority of the proposed hashing method over competitive hashing methods for different cross-modal retrieval tasks.

2. PROBLEM FORMULATION

In the remainder of the paper, we use the following notation, numbers are denoted by lower case letters e.g., a; matrices by plain upper case letters e.g., A; sets by calligraphic upper case letters e.g., $\mathcal{A}$; and vectors by lower case bold letters e.g., $\mathbf{a}$.

Let n be the number of documents and m the number of modalities. Given $j = 1 \dots m$ and $i = 1 \dots n$, each i-th document is represented in the j-th modality by a d_j-dimensional feature vector:

$$\mathbf{x}_i^{(j)} = [x_1^{(j)}, x_2^{(j)}, \dots, x_{d_j}^{(j)}] \in \Re^{1 \times d_j}$$

with $d_j \neq d_{j'}$ for two different modalities j and j'. In the j-th modality, the n feature vectors are stored in a matrix $X^{(j)} \in \Re^{d_j \times n}$, where the i-th column is the feature vector $\mathbf{x}_i^{(j)}$. In our method, we assume that the inter-modality and intra-modality similarities are calculated in a unified hash code with length k for each document i:

$$\mathbf{b}_i = [y_1, y_2, \dots, y_k] \in \{0,1\}^{1 \times k}$$

The problems that our cross-modal hashing method faces are formally defined as follows:

DEFINITION 1. *(Internal Documents' Hash Codes) "To compute a consensus binary matrix $B \in \{0,1\}^{k \times n}$, where the i-th column corresponds to the unified hash code $\mathbf{b}_i$ of an internal document i."*

DEFINITION 2. *(External Documents' Hash Codes) "To calculate m projection matrices $P^{(j)} \in \Re^{k \times d_j}$, with $j = 1, \dots, m$, so as for any external document i that does not belong to B, given its representation $\mathbf{x}_i^{(j)}$ in modality j to generate the respective hash code $\mathbf{b}_i$."*

In the following Section, we detail each step of our method.

3. PROPOSED APPROACH

3.1 Cross-Modal Cluster Representations

$\forall\ j{=}1, \dots, m$ we cluster the n feature vectors $\mathbf{x}^{(j)}$ with power iteration [7]. Thus, we have m different clusterings:

$$\begin{cases} \mathcal{C}_1^{(1)}, \mathcal{C}_2^{(1)}, \dots, \mathcal{C}_{q_1}^{(1)} \\ \mathcal{C}_1^{(2)}, \mathcal{C}_2^{(2)}, \dots, \mathcal{C}_{q_2}^{(2)} \\ \qquad \vdots \\ \mathcal{C}_1^{(m)}, \mathcal{C}_2^{(m)}, \dots, \mathcal{C}_{q_m}^{(m)} \end{cases} \tag{1}$$

where q_j is the number of clusters in the j-th modality, with $q_j \neq q_{j'}$ for two modalities j, j'. Also, $\mathcal{C}_w^{(j)}$ denotes the w-th cluster of the j-th modality, with $w \in 1, \dots q_j^{(j)}$. Let $Z \in \Re^{q \times n}$ be the cross-modal cluster matrix, with $q{=}q_1 + q_2{+}, \dots, q_m$. $\forall\ c{=}1, \dots, q$ we set $Z(c,i){=}1$, if a document i belongs to cluster c, and 0 otherwise. The i-th column of Z is a representation (feature vector) of document i based on all the generated clusters in the m modalities. Hence, if two documents, e.g. i and p, belong to the same clusters in most of the m modalities, then the respective i-th and p-th columns (representations) of the cross-modal cluster matrix Z will be similar.

3.2 Joint Matrix Factorization

3.2.1 Intra-modality and Cluster-based Similarities in Individual Matrix Factorization

To capture the intra-modality similarities of the n documents, $\forall\ j{=}1, \dots, m$ we consider the following matrix factorization of $X(j)$:

$$X^{(j)} \approx U^{(j)} V^{(j)} \tag{2}$$

with $U^{(j)} \in \Re^{d_j \times k}$, $V^{(j)} \in \Re^{k \times n}$ and k being the number of latent factors in matrix factorization, equal to the length of the hash codes $\mathbf{b_i}$. In addition, we consider the matrix factorization of the cross-modal cluster matrix Z as follows: $Z \approx U_z V_z$, with $U_z \in \Re^{q \times k}$ and $V^{(j)} \in \Re^{k \times n}$, to capture the associations of the n documents at the cluster-based level. Given the m matrices $X^{(j)}$ and the cross-modal cluster matrix Z, in total we have $m+1$ individual matrix factorizations:

$$\begin{cases} X^{(1)} \approx U^{(1)} V^{(1)} \\ X^{(2)} \approx U^{(2)} V^{(2)} \\ \qquad \vdots \\ X^{(m)} \approx U^{(m)} V^{(m)} \\ Z \approx U_z V_z \end{cases} \tag{3}$$

As presented in Section 3.1, the cluster assignments are non-negative ($Z \geq 0$), with the feature vectors also being non-negative in each matrix $X^{(j)} \geq 0$; consequently, each matrix factorization in Eq. (3) is subject to:

$$U^{(1)}, \dots U^{(m)} \geq 0, \quad V^{(1)}, \dots V^{(m)} \geq 0, \quad U_z, V_z \geq 0$$

3.2.2 Inter-modality, Intra-modality and Cluster-based Similarities in Joint Matrix Factorization

All matrices in Eq. (3) have to be jointly factorized to simultaneously compute the inter-modality and intra-modality similarities of the documents in the m modalities, and to capture the associations with the cross-modal cluster matrix. We formulate a joint matrix factorization process with the constraint that pushes the matrices $X^{(j)}$ of the m modalities and the cross-modal cluster matrix Z into a common consensus matrix $B^* \in \Re^{k \times n}$, with B^* corresponding to a shared k-dimensional latent space of the n internal documents. To consider the case of having the representation of an external document in modality j, we denote the projection of each modality j into the shared k-dimensional space as a projection matrix $P^{(j)} \in \Re^{k \times d_j}$ on condition that:

$$V^{(j)} = P^{(j)} X^{(j)}, \quad \text{with} \quad j = 1, \dots, m \tag{4}$$

where $V^{(j)} \in \Re^{k \times n}$ and $X^{(j)} \in \Re^{d_j \times n}$. To calculate the m projection matrices $P^{(j)}$ and the consensus matrix B^*, we formulate the problem of joint matrix factorization as a minimization problem [3]. Based on Eqs. (3) and (4), we have to minimize the following loss function $\mathfrak{L}$ for the joint matrix factorization:

$$\mathfrak{L}(U^{(1)}, \ldots, U^{(m)}, P^{(1)}, \ldots, P^{(m)}, U_z, V_z, B^*) = \sum_{j=1}^{m} \left\{ ||X^{(j)} - U^{(j)} P^{(j)} X^{(j)}||_F^2 + \lambda_j ||P^{(j)} X^{(j)} - B^*||_F^2 \right\} + ||Z - U_z V_z||_F^2 + \lambda_z ||V_z - B^*||_F^2 \quad (5)$$

where $||\cdot||_F$ denotes the Frobenius norm. In the second line of Eq. (5), the term $||X^{(j)} - U^{(j)} P^{(j)} X^{(j)}||_F^2$ denotes the approximation error of each factorized matrix $X^{(j)}$, while the term $||P^{(j)} X^{(j)} - B^*||_F^2$ is the *disagreement measurement* between the product $P^{(j)} X^{(j)}$ and the consensus matrix B^*. Accordingly, in the third line of Eq. (5), the first term $||Z - U_z V_z||_F^2$ denotes the approximation error of the factorized cross-modal cluster matrix Z; and $||V_z - B^*||_F^2$ is the respective *disagreement measurement*. The product $P^{(j)} X^{(j)}$, matrix V_z and the consensus matrix B^* are comparable, as they have the same dimensionality ($k \times n$). $\forall\ j = 1, \ldots, m$ each parameter λ_j tunes the weight of modality j over the joint factorization, and the respective disagreement term; accordingly, parameter λ_z for the cross-modal cluster matrix. For reasons of simplicity we set $\lambda_1 = \ldots = \lambda_m = \lambda_z = \lambda$. In our approach, we solve the minimization problem of $\mathfrak{L}$, using the "multiplicative rules" [3], an iterative update procedure, where in each iteration we fix B^* and minimize $\mathfrak{L}$ over the rest of matrices, and then we fix the rest of matrices and minimize $\mathfrak{L}$ over B^*; The outcome of the joint factorization when minimizing the loss function $\mathfrak{L}$ of Eq. (5) is computing matrices $U^{(1)}, \ldots, U^{(m)}, P^{(1)}, \ldots, P^{(m)}, U_z, V_z, B^*$.

3.3 Hash Codes

We generate the unified hash codes in $B \in \{0, 1\}^{k \times n}$ by binarizing the real values of the consensus matrix $B^* \in \Re^{k \times n}$, thus generating a binary vector $\mathbf{b}_i = [y_1, y_2, \ldots, y_k]$ for each internal document i. According to [8], efficient hash codes should also maximize the entropy; based on the maximum entropy principle, a binary bit that gives balanced partitioning of the whole data set should provide maximum information. Hence, given the code length k, then $\forall a = 1, \ldots, k$ we set a threshold $thres_a$ for binarizing the a-th bit, with $thres_a$ being the median[1] of the a-th row in B^*. If a bit $y_a > thres_a$, then we set $y_a = 1$ and 0 otherwise. In doing so, the binary code achieves the best balance. To generate the hash code of an external document i that does not belong to matrix B, we use the projection matrices, computed by the joint matrix factorization when minimizing the loss function in Eq (5). Given the feature vector $\mathbf{x}_i^{(j)} \in \Re^{1 \times d_j}$ of the external document i in the j-th modality, we use the respective projection matrix $P^{(j)} \in \Re^{k \times d_j}$ to calculate the following real-value vector:

$$\mathbf{b}_i = \mathbf{x}_i^{(j)} P^{(j)^T} \in \Re^{1 \times k} \quad (6)$$

which is then binarized as in the internal case, creating the hash code $\mathbf{b}_i$ for the external document i.

[1]We use the median as a threshold, because mean values are sensitive to extremely high or low values in B^*.

4. EXPERIMENTAL EVALUATION

4.1 Data Sets

We use two real-world data sets *Wiki*[2] and *NUS-WIDE*[3]. The *Wiki* data set consists of 2,866 Wikipedia documents, with each document containing a text and one corresponding relevant image. In addition, images are represented by 128-dimensional SIFT feature vectors [10] and text by 10-dimensional topics vectors. All documents (image-text pairs) are labeled by 10 semantic categories. We split the data set as follows, 20% as query set; 5% as cross-validation set to tune the parameters of each method; while the remaining 75% of the data set is considered as the database, from which the hash codes are learnt and the results are retrieved. *NUS-WIDE* consists of 269,648 image-tag pairs from Flickr, where we keep the image-tag pairs that belong to one of the 10 largest concepts [11, 14]. Images are represented by 500-dimensional SIFT vectors and text as 1000-dimensional vectors, by performing PCA on the original tag occurrences [14]. The query set consists of 1% of the dataset; 1% is used as cross-validation set; while the remaining data set is the database to retrieve the results. The hash codes are learned from 5K image-text pairs from the database [2].

4.2 Evaluation Protocol

As both data sets are bi-modal with images and texts, we evaluate our hashing method on the following cross-modal retrieval tasks: (i) image-query → text results and (ii) text-query → image results. Following the evaluation protocol of relevant studies [2, 11, 14], we measure the performance of cross-modal hashing in terms of mAP. Given a query and the top result set $\mathcal{R}$, Average Precision (AP) is defined as:

$$AP = \frac{1}{l} \sum_{r \in \mathcal{R}} P(r)\delta(r) \quad (7)$$

where l is the number of true neighbors in the retrieved set $\mathcal{R}$, that is, the results which belong to the same category with the query; $P(r)$ denotes the precision of the top retrieved results in set $\mathcal{R}$ and $\delta(r)$=1 if the r-th result is a true neighbor and 0 otherwise. In the experiments we set $|\mathcal{R}|$=50, and mAP is computed by averaging the AP values over all the queries. We repeated our experiments five times and we report mean mAP values and standard deviations over the runs.

4.3 Results

In the proposed **C**luster-based **J**oint **M**atrix **F**actorization **H**ashing method (**C-JMFH**), we varied the parameter λ in [10^{-4} 10^{-1}], concluding in 10^{-2} and 10^{-3} for *Wiki* and *NUS-WIDE*, respectively. To evaluate the impact of the cross-modal cluster matrix when computing the consensus matrix B^* in Eq. (5), and consequently the binary codes in B, we use as baseline a variant of our method, namely **JMFH**, which does not consider the cross-modal cluster matrix in the joint factorization process. We use **IMH**[4] [12] as baseline, and we compare the proposed C-JMFH method

[2]http://www.svcl.ucsd.edu/projects/crossmodal/
[3]http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm
[4]http://staff.itee.uq.edu.au/shenht/publications.htm

with **CMFH**[5] [2], a hashing method that also follows a joint matrix factorization strategy when generating the binary codes. The parameter values in the baselines were determined by cross-validation, using the publicly available implementations and we report the best results.

Figures 1 and 2 show the results for the cross-modal retrieval tasks in *Wiki* and *NUS-WIDE*, respectively. The baseline IMI method has poor performance, compared with the rest of cross-modal hashing methods, by not following a joint matrix factorization strategy and thus not computing the inter-modality and intra-modality similarities as well as CMFH, JMFH and C-JMFH do. As also observed in relevant studies [2, 6], the retrieval accuracy does not necessarily improve when increasing the number of bits in many baseline methods such as IMI, as it does not follow a joint matrix factorization strategy. Meanwhile, we observe that CMFH and JMFH achieve comparable performance, as both hashing methods jointly factorize the representations in the different modalities, with the retrieval accuracy increasing when a larger number of bits is selected. This happens because in these methods the hamming space corresponds to the latent space of the joint matrix factorization process, consequently encoding more information for a large number of bits. The proposed C-JMFH method boosts the *mAP* retrieval accuracy by incorporating the cross-modal cluster representations of the documents when learning the hash codes. Compared to the second best method, the proposed C-JMFH method achieves a relative improvement of 3.1-12.9% in all the cross-modal retrieval tasks, with the exceptional case of 1.9% relative improvement in the case of a low selection of number of bits, that is, the case of 16 bits in *NUS-WIDE* for the text → image cross-modal retrieval task. To verify the superiority of the proposed C-JMFH method, we used the paired t-test and we found that the differences between the reported results for C-JMFH against the competitive hashing methods were statistically significant for $p<0.05$.

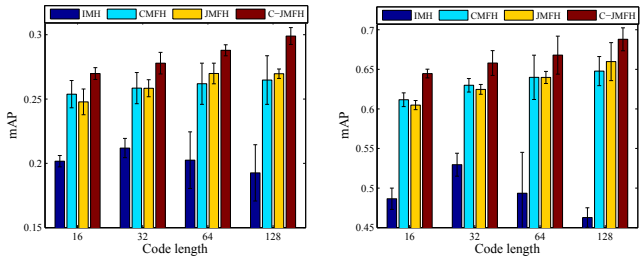

Figure 1: Performance evaluation on *Wiki* for the cross-modal retrieval tasks.

5. CONCLUSION

In this study, we presented a cross-modal hashing method using a cluster-based representation into a joint factorization process. In our experiments, we showed that the proposed method outperforms other state-of-the-art hashing methods in terms of cross-modal retrieval accuracy. Although we focused on the case of cross-modal retrieval, that is, given a query from one modality to retrieve results from another modality, an interesting future direction is to extend our method to multimodal retrieval, where given a query from one or more modalities to retrieve documents from all the different modalities [9].

[5]http://ise.thss.tsinghua.edu.cn/MIG/publications.jsp

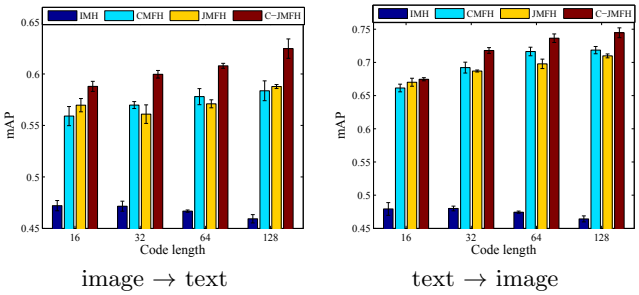

Figure 2: Performance evaluation on *NUS-WIDE*.

6. REFERENCES

[1] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*, pages 3594–3601, 2010.

[2] G. Ding, Y. Guo, and J. Zhou. Collective matrix factorization hashing for multimodal data. In *CVPR*, pages 2083–2090, 2014.

[3] J. Gao, J. Han, J. Liu, and C. Wang. Multi-view clustering via joint nonnegative matrix factorization. In *SDM*, pages 252–260, 2013.

[4] A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In *VLDB*, pages 518–529, 1999.

[5] Y. Gong and S. Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *CVPR*, pages 817–824, 2011.

[6] S. Kumar and R. Udupa. Learning hash functions for cross-view similarity search. In *IJCAI*, pages 1360–1365, 2011.

[7] F. Lin and W. W. Cohen. Power iteration clustering. In *ICML*, pages 655–662, 2010.

[8] R. Lin, D. A. Ross, and J. Yagnik. SPEC hashing: Similarity preserving algorithm for entropy-based coding. In *CVPR*, pages 848–854, 2010.

[9] X. Liu, Y. Mu, B. Lang, and S. Chang. Mixed image-keyword query adaptive hashing over multilabel images. *TOMCCAP*, 10(2):22, 2014.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[11] S. Moran and V. Lavrenko. Regularised cross-modal hashing. In *SIGIR*, pages 907–910, 2015.

[12] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *SIGMOD*, pages 785–796, 2013.

[13] Y. Wang, X. Lin, L. Wu, W. Zhang, and Q. Zhang. Lbmch: Learning bridging mapping for cross-modal hashing. In *SIGIR*, pages 999–1002, 2015.

[14] Y. Zhen and D. Yeung. Co-regularized hashing for multimodal data. In *NIPS*, pages 1385–1393, 2012.