MODALITY-SPECIFIC STRUCTURE PRESERVING HASHING FOR CROSS-MODAL RETRIEVAL

Xingbo Liu[1], *Xiushan Nie*[1,2] *, *Haoliang Sun*[1], *Chaoran Cui*[2], *Yilong Yin* [1*]

[1]School of Computer Science and Technology, Shandong University, Jinan, P.R. China
[2]School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, P.R. China
sclxb@mail.sdu.edu.cn,{niexsh,crcui}@sdufe.edu.cn , haolsun.cn@gmail.com,ylyin@sdu.edu.cn

ABSTRACT

Hashing-based methods have made great advancements in cross-modal retrieval in both computational efficiency and storage. Learning a common space from different modalities is the common strategy of hashing-based methods, however, relational and structural information between samples in each modality, namely, a modality-specific structure, is always discarded during learning. In addition, cross-modality samples sometimes suffer from inter-class ambiguity and intra-class variability because of the uncertainty of manual labeling. To address these issues, we propose a novel method named Modality-specific structure Preserving Hashing (MsPH), which learns hashes by preserving the local structure and relations between samples in each modality. Moreover, label enhancement is utilized in MsPH to address label ambiguity and variability. Extensive experiments conducted on three benchmark datasets demonstrate the superiority of MsPH under various cross-modal scenarios.

Index Terms— Cross-modal retrieval, Hashing, Modality-specific structure preserving, Label enhancement

1. INTRODUCTION

With the rapid development of the internet, a vast amount of media data such as text, image, and video have been captured and shared on social media sites such as Flickr and Facebook. When users query topics with texts, they often expect relevant images or videos with similar semantics to be returned; this necessitates the development of cross-modal retrieval technologies to meet latent commercial needs. Cross-modal retrieval aims to retrieve information from different modalities such as images, text, or videos and has become a hot topic in the field of multimedia retrieval. In general, an advantageous solution to cross-modal retrieval is hashing-based methods, which compress high-dimensional data into compact binary codes with similar binary codes for similar objects. Hashing-based methods provide both computational efficiency and search quality.

Many cross-modal hashing methods have been proposed in recent years. Unsupervised [1] [2] [3] and supervised [4] [5] [6] [7] methods are two main categories. Unsupervised hashing methods generally aim to learn projections from features to hash codes by exploiting their intra- and inter-view relationships of training data. Song et al. [1] propose a novel inter-media hashing model to derive effective hash codes by exploring inter- and intra-media consistencies. Zhou et al. [2] propose latent semantic sparse hashing to perform a cross-modal similarity search using sparse coding and matrix factorization to obtain latent semantics. Ding et al. [3] learn unified hash codes by collective matrix factorization with a latent factor model from different modalities of one instance. Supervised hashing-based methods leverage available class label information of training data as supervision to facilitate hash code learning. Zhang et al. [4] utilize label vectors to obtain semantic similarity matrices and attempted to reconstruct them through learned hash codes. Lin et al. [5] transform given semantic affinities of training data into a probability distribution and approximated them with to-be-learned hash codes in Hamming space.

In general, most existing methods attempt to effectively reflect the corresponding relation of different modalities with semantic meanings or label information during hash learning. However, local structure and relation information in each modality, namely, a modality-specific structure, is always discarded when learning a common space for different modalities. Different modalities lie on different manifolds, and the modality-specific structure obviously benefits the process of hashing learning. Therefore, it is important to preserve the modality-specific structure in the final hash representation. However, few researchers have taken this into consideration. In addition, because samples from different modalities always share identical and abstract labels that are manually labeled, which is imprecise and subject to error, cross-modality data often suffer from inter-class ambiguity and intra-class variability.

Xiushan Nie and Yilong Yin are both the corresponding authors of this work.

To address these issues, we propose a novel hashing-based method for cross-modality retrieval named Modality-specific structure Preserving Hashing (MsPH), which not only determines a common space for different modalities, but also preserves the local manifold structure of each modality. We utilize a label enhancement strategy to release the inter-class ambiguity and intra-class variability during the retrieval process. The main contributions of this work are two-fold:

- *Modality-specific structure preservation*. The local structures and relations between different samples in each modality are preserved and considered in hash learning to protect every modality distribution or geometrical structure.
- *Label enhancement*. Label enhancement is considered during cross-modal retrieval, which not only releases the inter-class ambiguity and intra-class variability, but also satisfies user-central and content-central retrieval.

2. THE PROPOSED METHOD

This section details the proposed hashing method for cross-modal retrieval. For simplicity, and without loss of generality, we only discuss the case of two modalities, i.e., image and text. Cases that consider more modalities can easily be extended using the proposed method.

The framework of MsPH is illustrated in Fig. 1. MsPH attempts to learn the hash representation from different modalities by considering the local structure and relation of samples in each modality. It then retrieves relevant samples based on hashes and label enhancement. In this section, we describe hash learning and the label enhancement-based retrieval process.

2.1. Hash Learning

Assume that we have a training set $\mathbf{X}$ consisting of n instances, i.e., $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ with $\mathbf{x}_i$ being the i_{th} instance. For each instance, $\mathbf{x}_i = (\mathbf{a}_i, \mathbf{b}_i)$, where $\mathbf{a}_i \in R^f$is the image and $\mathbf{b}_i \in R^d$ is the text feature vectors, respectively. Moreover, label matrix $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n$ is also available with $\mathbf{y}_i = \{y_{ij}\} \in R^c$ being the label vector of the i_{th} instance, where c is the total number of categories, and $y_{im} = 1$ if $\mathbf{x}_i$ belongs to class m, $m = 1, \cdots, c$ and 0 otherwise. We also define $\mathbf{A} \in R^{f\times n}$ and $\mathbf{B} \in R^{d\times n}$ as the image and text feature matrixcs, respectively. Then, the goal of MsPH is to learn hash matrix $\mathbf{H}$ for training instances.

2.1.1. Formulation

The goal of hash learning is to obtain a common hash representation shared by different modalities. First, we define $\mathbf{U}_1$ and $\mathbf{U}_2$ as transformation matrices that map the image and text features to the hashing space, respectively. In order to effectively utilize label information, transformation matrix $\mathbf{M}$ is utilized to map hash matrix $\mathbf{H}$ to its label space. To learn hash codes from different modalities, minimizing the errors of both the projection and classification procedures is a reasonable criterion that can be applied; this can be formulated as the following minimization problem:

$$\begin{aligned} &\min_{\mathbf{M},\mathbf{H},\mathbf{U_1},\mathbf{U_2}} \left\|\mathbf{Y}-\mathbf{M}^T\mathbf{H}\right\|^2 + u_1\left\|\mathbf{H}-\mathbf{U}_1^T\mathbf{A}\right\|^2 \\ &+ u_2\left\|\mathbf{H}-\mathbf{U}_2^T\mathbf{B}\right\|^2 + \lambda\left\|\mathbf{M}\right\|^2, \\ &s.t. \quad \mathbf{H} \in \{-1,1\}^{K\times n}, \end{aligned} \tag{1}$$

where u_1 and u_2 are penalty parameters, λ is a regularization parameter, $\|\cdot\|$ is the ℓ_2 norm and K is the length of hash codes .

Clearly, different modalities always lie on different manifolds that describe their modality-specific distributions. However, when samples in different modalities are projected into a common Hamming space, the modality-specific distributions may be discarded. Therefore, to obtain the optimized hash representation, the local structure in each modality must be preserved in the common space. To be specific, for $\mathbf{S} \in R^{n\times n}$, $S_{i,j} = 1$ if the i_{th} instance shares at least one label with the j_{th} instance, and $S_{i,j} = 0$ otherwise. We then derive the corresponding normalized Laplacian matrix as follows:

$$L(\mathbf{S}) = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\,\mathbf{S}\,\mathbf{D}^{-\frac{1}{2}}, \tag{2}$$

where $\mathbf{I} \in R^{n\times n}$ is the identity matrix, and $\mathbf{D} \in R^{n\times n}$ is the diagonal degree matrix whose $(u,u)_{th}$ entry is the sum of the u_{th} row of $\mathbf{S}$.

Our goal is to preserve the local structure in each modality when they are projected onto a common Hamming space. In this study, we minimize structure regularization items $LP_{\mathbf{A}}$ and $LP_{\mathbf{B}}$ for image and text modalities, respectively, to achieve this purpose. $LP_{\mathbf{A}}$ and $LP_{\mathbf{B}}$ are defined as follows:

$$LP_{\mathbf{A}} = Tr((\mathbf{U}_1^T\mathbf{A})^T L(\mathbf{S})(\mathbf{U}_1^T\mathbf{A})), \tag{3}$$

$$LP_{\mathbf{B}} = Tr((\mathbf{U}_2^T\mathbf{B})^T L(\mathbf{S})(\mathbf{U}_2^T\mathbf{B})), \tag{4}$$

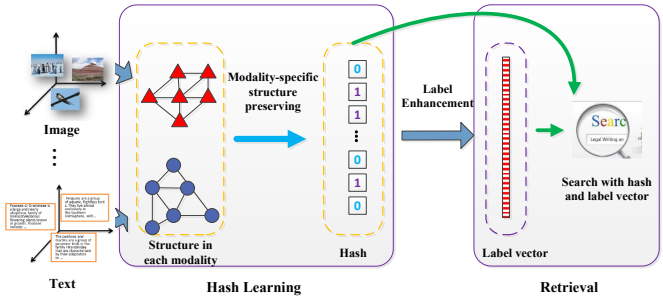

Fig. 1. The framework of the modality-specific structure preserving hashing(MsPH).

where $Tr(\cdot)$ is the trace.

Eventually, the objective function for MsPH can be concluded as:

$$\begin{aligned}&\min_{\mathbf{M},\mathbf{H},\mathbf{U_1},\mathbf{U_2}} \left\|\mathbf{Y}-\mathbf{M}^T\mathbf{H}\right\|^2 + u_1\left\|\mathbf{H}-\mathbf{U}_1^T\mathbf{A}\right\|^2 \\ &+ u_2\left\|\mathbf{H}-\mathbf{U}_2^T\mathbf{B}\right\|^2 + \alpha * LP_{\mathbf{A}} + \beta * LP_{\mathbf{B}} + \lambda\left\|\mathbf{M}\right\|^2, \\ &s.t. \quad \mathbf{H}\in\{-1,1\}^{K\times n},\end{aligned} \tag{5}$$

where α and β are regularization parameters.

2.1.2. Optimization

It's challenging to optimize Eq.(5) directly as it is non-convex. However, it is convex when taking one variable with the other three variables fixed. Therefore, Eq.(5) can be solved by an iterative framework with the following steps until convergence. In the Eq.(5), the variables $\mathbf{M}$, $\mathbf{H}$, $\mathbf{U}_1$ and $\mathbf{U}_2$ are the ones we need to optimize, while $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{Y}$ are what we know. In addition, the variables $\mathbf{M}$ and $\mathbf{H}$ depend on $\mathbf{Y}$, while $\mathbf{U}_1$ and $\mathbf{U}_2$ depend on $\mathbf{H}$. As far as the optimized order is concerned, it is better to optimize $\mathbf{M}$ and $\mathbf{H}$ firstly. Therefore, Eq.(5) can be solved as three steps.

Step1: Learn M with the other variables fixed. The problem in Eq.(5) becomes:

$$\min_{\mathbf{M}} \left\|\mathbf{Y}-\mathbf{M}^T\mathbf{H}\right\|^2 + \lambda\left\|\mathbf{M}\right\|^2, \tag{6}$$

which is a simple regularized least squares problem. The closed-form solution of $\mathbf{M}$ can be derived as:

$$\mathbf{M} = (\mathbf{H}\mathbf{H}^T+\lambda)^{-1} * (\mathbf{H}^T\mathbf{Y}). \tag{7}$$

Step2: Learn the binary code in common space $\mathbf{H}$ with the other variables fixed. The problem in Eq.(5) becomes:

$$\begin{aligned}&\min_{\mathbf{H}} \left\|\mathbf{Y}-\mathbf{M}^T\mathbf{H}\right\|^2 + u_1\left\|\mathbf{H}-\mathbf{U}_1^T\mathbf{A}\right\| \\ &+ u_2\left\|\mathbf{H}-\mathbf{U}_2^T\mathbf{B}\right\|^2, \qquad s.t. \quad \mathbf{H}\in\{-1,1\}^{K\times n}.\end{aligned} \tag{8}$$

We rewrite Eq.(8) as:

$$\min_{\mathbf{H}} \left\|\mathbf{M}^T\mathbf{H}\right\|^2 - Tr(\mathbf{H}^T\mathbf{W}), \qquad s.t. \quad \mathbf{H}\in\{-1,1\}^{K\times n}, \tag{9}$$

where $\mathbf{W} = \mathbf{M}\mathbf{Y} + u_1\mathbf{U}_1^T\mathbf{A} + u_2\mathbf{U}_2^T\mathbf{B}$, and $Tr(\cdot)$ is the trace.

Although Eq.(9) is a NP hard problem as $\mathbf{H}$ is discrete, we can directly leverage the discrete cyclic coordinate descent (DCC) approach proposed in [11] to learn $\mathbf{H}$ bit-by-bit iteratively. Specifically, define $\mathbf{h}^T$ as the k_{th} row of matrix $\mathbf{H}$, $k = 1, \cdots, K$ and $\mathbf{H}'$ as the matrix $\mathbf{H}$ excluding $\mathbf{h}$. Homoplastically, define $\mathbf{v}^T$ as the k_{th} row of matrix $\mathbf{M}$ and $\mathbf{M}'$ as the matrix $\mathbf{M}$ excluding $\mathbf{v}$. Besides, define $\mathbf{w}^T$ as the k_{th} row of matrix $\mathbf{W}$. Then optimal solution of can be achieved as:

$$\mathbf{h} = sgn(\mathbf{w} - \mathbf{H}'^{\mathbf{T}}\mathbf{M}'\mathbf{v}). \tag{10}$$

DCC can get optimal solution of the binary code of $\mathbf{H}$ bit-by-bit, by which every bit of the $\mathbf{H}$ has be computed depends on the $K-1$ learned bits. After K iterations, the matrix $\mathbf{H}$ is supposed to be optimal for Eq.(8).

Step3: Respectively learn the projection matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ with other variables fixed. The problem in Eq.(5) becomes:

$$\min_{\mathbf{U}_1} \alpha * LP_{\mathbf{A}} + u_1\left\|\mathbf{H}-\mathbf{U}_1^T\mathbf{A}\right\|^2, \tag{11}$$

$$\min_{\mathbf{U}_2} \beta * LP_{\mathbf{B}} + u_2\left\|\mathbf{H}-\mathbf{U}_2^T\mathbf{B}\right\|^2. \tag{12}$$

The two matrices can be computed by partial derivative as:

$$\mathbf{U}_1 = (\alpha * \mathbf{A}^T L(\mathbf{S})^T\mathbf{A} + \mathbf{A}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{H}^T), \tag{13}$$

$$\mathbf{U}_2 = (\beta * \mathbf{B}^T L(\mathbf{S})^T\mathbf{B} + \mathbf{B}\mathbf{B}^T)^{-1}(\mathbf{B}\mathbf{H}^T). \tag{14}$$

In short, Eq.(5) can be solved iteratively by above three steps. The convergence is supposed to be reached in three to five times of iteration.

2.2. Label Enhancement-based Retrieval

In reality, multiple modality samples are manually labeled with brief labels. Sometimes samples from different classes likely share similar content, while samples belonging to the same class differ heavily in content. So these labels are not accurate. For example, all kinds of animals belong to the class biology, while photos of humans may belong to history, sports, music, or media. To tackle this problem, label enhancement is utilized in this study. Label enhancement [12] describes every sample using the probability of assigning the sample to different labels.

In this study, we utilize transformation matrix $\mathbf{M}$ to achieve label enhancement. As discussed above, matrix $\mathbf{M}$ is obtained by MsPH and utilized to project the common hash representation to the label space. Based on $\mathbf{M}$, a label vector $\widehat{\mathbf{y}}$ that is called the label distribution for a sample can be obtained, where the k_{th} element represents the probability of assigning the sample to the k_{th} label. Compared to existing methods that only use an identical label for a sample, label enhancement can release the inter-class ambiguity and intra-class variability.

During retrieval, we consider both the label distribution and hash representation. Unlike existing methods, we not only utilize the hash distance in the common space to decide whether a retrieved sample is relevant, but we also consider the similarity of label distribution. In general, two strategies are adopted: 1) discarding samples whose labels are very different from the query's label after sorting the retrieved samples by hash distance, and 2) establishing a simple weighting on the hash distance of contents and the similarity of label distributions.

Table 1. Overall comparison of *mAP* values on the three datasets. The top panel is the performance for Img2Text task and the bottom panel is for Text2Img task.

Method	Wiki				MIR-flickr				NUS-WIDE			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
CMSSH [8]	0.1976	0.1999	0.1889	0.1907	0.5520	0.5539	0.5506	0.5559	0.4686	0.4768	0.4741	0.4637
IMH [1]	0.1869	0.1938	0.1873	0.1834	0.6088	0.6063	0.5977	0.5857	0.4543	0.4469	0.4371	0.4285
LSSH [2]	0.2141	0.2216	0.2218	0.2211	0.5784	0.5804	0.5797	0.5816	0.3900	0.3924	0.3962	0.3966
CMFH [3]	0.2132	0.2259	0.2362	0.2419	0.6273	0.6343	0.6410	0.6451	0.4267	0.4229	0.4207	0.4182
CRH [9]	0.2031	0.1966	0.1982	0.1943	0.5826	0.5745	0.5726	0.5718	0.5136	0.5079	0.4996	0.5013
DCH [10]	0.2578	0.2791	0.2926	0.2661	0.5268	0.5192	0.5079	0.6797	0.4859	0.5811	0.5403	0.5571
MsPH	0.2771	0.2955	0.2980	0.2896	0.6960	0.7216	0.7370	0.7456	0.5722	0.6047	0.5588	0.5717
CMSSH [8]	0.2495	0.2360	0.2348	0.2382	0.6010	0.6048	0.6029	0.6041	0.4635	0.4685	0.4594	0.4556
IMH [1]	0.3731	0.3967	0.3720	0.3519	0.5996	0.5999	0.5914	0.5824	0.4546	0.4497	0.4415	0.4304
LSSH [2]	0.5031	0.5224	0.5293	0.5346	0.5898	0.5927	0.5932	0.5932	0.4286	0.4248	0.4248	0.4175
CMFH [3]	0.4884	0.5132	0.5269	0.5375	0.6095	0.6134	0.6184	0.6199	0.4627	0.4556	0.4518	0.4478
CRH [9]	0.2634	0.2622	0.2631	0.2625	0.5944	0.5913	0.5838	0.5811	0.5273	0.5114	0.5033	0.4977
DCH [10]	0.3801	0.4237	0.4431	0.4049	0.6109	0.6407	0.6221	0.6576	0.5984	0.5993	0.5852	0.6103
MsPH	0.4563	0.4670	0.4724	0.4709	0.7296	0.7355	0.7354	0.7613	0.6142	0.6403	0.6256	0.6117

3. EXPERIMENT

3.1. Experimental Settings

To confirm the superiority of our method, we conduct sufficient experiments on three benchmark datasets, Wiki [13], MIR-flickr [14], and NUS-WIDE [15]. All of these datasets consist of pairwise parts, images, and texts. Two types of cross-modal retrieval tasks are conducted on the three benchmark datasets: 1) Img2Text: using images to query related texts, and 2) Text2Img: using texts to query related images.

We compare the proposed MsPH with state-of-the-art hashing based cross-modal retrieval methods such as Cross-modal Similarity Sensitive Hashing (CMSSH) [8], Intermedia Hashing (IMH) [1], Latent Semantic Sparse Hashing (LSSH) [2], Collective Matrix Factorization Hashing (CMFH) [3], Co-Regularized Hashing (CRH) [9], and Discrete Cross-modal Hashing (DCH) [10]. We adopt the *mean average precision* (*mAP*) to evaluate the performance of these methods on the three public datasets.

To verify the stability of MsPH, we perform five runs for our methods and average their performance for comparison. For the experimental parameters, we empirically set $\lambda = 10^{-2}$, $\alpha = 0.4$, $\beta = 0.5$, and $u_1 = u_2 = 10^{-6}$. During the retrieval, we empirically set the weights of hash distance and label similarity to 0.8 and 0.2, respectively. All experiments are conducted on a computer with an Intel Core i7-6700 3.40 GHz 4 processor and 16 GB RAM. The operating system is 64-bit Windows 10, and the programming environment is MATLAB R2015b.

3.2. Experimental Results and Analysis

Table 1 shows the *mAP* of each method with the hash length ranging from 16 to 128 bits using the same experimental settings defined in [10]. In the Wiki dataset, note that the *mAP* of MsPH is slightly higher than those of the other methods in the Img2Text task; similar performance can be observed in the Text2Img task. This is not immediately obvious because of the small size of this dataset. In contrast, on larger datasets MIR-flickr and NUS-WIDE, the advantage of the proposed MsPH is clearly evident. Particularly, in the MIR-flickr dataset, the *mAP* performance of MsPH achieved more than 10% improvement compared to the other methods both in the Img2Text and Text2Img tasks.

4. CONCLUSION

In this paper, we propose a supervised hashing method for cross-modal retrieval called MsPH, which focuses on modality-specific structure preservation and label enhancement during hashing learning and retrieval. In the proposed method, we consider a modality-specific structure regularization item while learning the common hashing space, making it more suitable for precise retrieval and cross-modal retrieval. Furthermore, we utilize label enhancement to release the inter-class ambiguity and intra-class variability in cross-modal samples. The experimental results conducted on three benchmark datasets confirm the superiority of our method.

5. ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (61671274, 61573219, 61701281), China Postdoctoral Science Foundation (2016M592190), Shandong Provincial Key Research and Development Plan (2017CXGC1504), Shandong Provincial Natural Science Foundation (ZR2017QF009), Shandong Provincial High College Science and Technology Plan (J17KB161), and the Fostering Project of Dominant Discipline and Talent Team of Shandong Province Higher Education Institutions.

6. REFERENCES

[1] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM, 2013, pp. 785–796.

[2] Jile Zhou, Guiguang Ding, and Yuchen Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014, pp. 415–424.

[3] Guiguang Ding, Yuchen Guo, and Jile Zhou, "Collective matrix factorization hashing for multimodal data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2075–2082.

[4] Dongqing Zhang and Wu-Jun Li, "Large-scale supervised multimodal hashing with semantic correlation maximization.," in *the Association for the Advance of Artificial Intelligence*, 2014, p. 7.

[5] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3864–3872.

[6] Xiushan Nie, Yilong Yin, Jiande Sun, Ju Liu, and Chaoran Cui, "Comprehensive feature-based robust video fingerprinting using tensor model," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 785–796, 2017.

[7] Xiushan Nie, Xiaoyu Li, Yane Chai, Chaoran Cui, Xiaoming Xi, and Yilong Yin, "Robust image fingerprinting based on feature point relationship mining," *IEEE Transactions on Information Forensics and Security*, 2018.

[8] Michael M Bronstein, Alexander M Bronstein, Fabrice Michel, and Nikos Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3594–3601.

[9] Yi Zhen and Dit-Yan Yeung, "Co-regularized hashing for multimodal data," in *Advances in neural information processing systems*, 2012, pp. 1376–1384.

[10] Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, and Xuelong Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2494–2507, 2017.

[11] Fumin Shen, Chunhua Shen, Wei Liu, and Heng Tao Shen, "Supervised discrete hashing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 37–45.

[12] Xin Geng, "Label distribution learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1734–1748, 2016.

[13] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 251–260.

[14] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[15] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng, "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*. ACM, 2009, p. 48.