ASYMMETRIC SCALABLE CROSS-MODAL HASHING

Wenyun Li, Chi-Man Pun

Department of Computer and Information Science
University of Macau
Macau, China

ABSTRACT

Cross-modal hashing is a practical approach to solving the problem of large-scale multimedia retrieval. However, there are still specific issues that the current methods cannot solve, such as how to construct binary codes rather than relax them to continuity effectively and how to prevent \( n \times n \) problem. This paper proposes a novel Asymmetric Scalable Cross-Modal Hashing (ASCMH) to address these issues. It learns a common latent space from the kernelized features of different modalities. It then transforms the similarity optimization to a distance-distance difference minimization problem with the help of semantic labels and common latent space. Additionally, we use an orthogonal constraint of label information to construct hash codes necessary for search accuracy. Extensive experiments on three benchmark datasets show that our ASCMH outperforms the SOTA cross-modal hashing methods.

Index Terms— Asymmetric hashing, Discrete optimization, Supervised hashing

1. INTRODUCTION

Multimedia data generation has increased as a result of Internet development. Multimedia retrieval and search might be severe issues. A valid and effective technique called cross-modal hashing tries to provide a hash code that maps highly dimensional data to a small, equivalent binary code. It has significantly focused on media retrieval to look for parallels in heterogeneous modalities. Generally speaking, we divide cross-modal hashing techniques into two classes: supervised and unsupervised cross-modal hashing.

Numerous unsupervised cross-modal hashing techniques exist. Such as CMFH[1], which uses collective matrix factorization to obtain hashing codes from the shared latent semantic space of cross-modalities, and DRMFH[2], which ideally makes use of a unified objection function to obtain consistency and inconsistency from various modalities. Therefore, supervised cross-modal hashing techniques benefit from semantic labelling. There are a ton of approaches that have been developed. Such as SMH[3], which can continuously learn the hash functions while being guided by label information; SCMH-WR[4], which can generate the binary code directly without relaxing binary constraints; and more recently, due to the popularity of deep learning. It is worth mentioning that with the success of deep neural networks, many deep learning-based cross-modal hashing methods[5] are proposed. Deep neural networks have a solid ability to fit the nonlinear representation and extract more discerning representation features. [6] includes features learning step and hashing generating step, and it takes advantage of the nonlinear modelling ability of deep learning network. [7] proposed a model which combines hashing encoding and classification and can easily discriminate the data pair and preserve the semantic information.

However, there are many issues with these techniques that must be taken into account: 1) Most supervised hashing techniques use label information’s inner product to create a semantic affinity matrix. Because they use \( n \times n \) pairwise similarity matrix, the memory and computing costs are reasonable for big datasets. 2) Because the hash code is a discrete optimization issue that is challenging to solve, many approaches relax the discrete hash code to a continuous one. However, this results in low-quality hash codes and high quantization errors. 3) Some hashing techniques simultaneously learn the hash function and codes, adding complexity.

We suggest Asymmetric Scalable Cross-Modal Hashing, a novel supervised cross-modal hashing technique, to simultaneously address these issues (ASCMH). The overview of our method is shown as Fig 1. To extract the shared semantic latent space from several modalities, we use collective matrix factorization. We use the distance-distance difference minimization method to maintain pairwise similarity while avoiding the enormous \( n \times n \) similarity matrix, dramatically reducing time and space complexity. We use the orthogonalization quantization between the semantic labels and the created hashing codes to fully leverage the label information. It is important to note that the hash codes in the suggested ASCMH are not produced via a relaxation process but separately. In order to increase the discriminating power, the...
ASCMH is a two-step method that can exchange out other potent models for the appropriate hashing function.

The main contributions of this paper are summarized as follows:

1. Asymmetric Scalable Cross-Modal Hashing (ASCMH) is used to present a unique supervised cross-modal hashing technique. It can change the discrete pairwise similarity matrix optimization into a problem of minimizing the distance between two points. As a result, large-scale datasets can use our method effectively.

2. In our ASCMH, a discrete optimization strategy for hashing codes is proposed. The high quantization error can be avoided by creating binary hash codes without relaxing.

3. Extensive tests are carried out on the three benchmark datasets Wiki, MIRFlickr-25K, and NUS-WIDE. The outcomes show that our approach outperforms several cutting-edge hashing techniques for cross-modal retrieval.

Fig. 1. The framework of the proposed Asymmetric Scalable Cross-Modal Hashing (ASCMH).

2. PROPOSED METHOD

Let $X^{(t)} = \{x_i^{(t)}\}_{i=1}^n$ be the training set, where $x_i^{(t)} \in \mathbb{R}^{d_t}$ will be defined as the $i$-th training data point from the $t$-th modality. In addition, all training data are centralized in our experiments. Moreover, in our model, $L = \{l_i\}_{i=1}^n \in \{0,1\}^{c \times n}$ is the label matrix. Our ASCMH’s goal is attaining the unified hashing codes from training data points and modality-specific hashing generation functions for cross-modal retrieval of testing data points.

2.1. The Proposed Model

First we adopt the collective matrix factorization technique [1] to extract the common semantic latent space from different modalities. So latent space learning is like the following:

$$
\min_{V,R,M} \sum_{i=0}^{l} \lambda \phi\left(X^{(r)} - P^{(r)}V\right)
$$

where $\sum_{r=0}^{l} \lambda_r = 1$ is the hyper-parameter for balance. $P^{(r)} \in \mathbb{R}^{k \times r}$ is the projection matrix, and $V \in \mathbb{R}^{r \times n}$ is the desired common semantic latent space. We transform similarity matrix by solving a distance-distance difference minimization problem as inspired by [8] and [9]. It means the distance difference between label pairwise and hash code pairwise shall be very small. So we have the following:

$$
\min_{B} \sum_{i,j=1}^{n} \left(\|B_i - B_j\|_F^2 - r\|G_i - G_j\|_F^2\right)
$$

$$
s.t. B \in \{-1, 1\}^{r \times n}
$$

where $G$ is a l2-norm normalized label matrix and defined as $G_i = L_i/\|L_i\|$. Then we turn Eqn. 2 as following:

$$
\min_{B} \sum_{i,j=1}^{n} \left(\|B_i - B_j\|_F^2 - r\|G_i - G_j\|_F^2\right)
$$

$$
\|B^T B - rG^TG\|_F^2, s.t.B \in \{-1, 1\}^{r \times n}
$$

In Eqn. 3, semantic information transforms to inner product semantic information search. Then we have learned the common semantic latent space $V$ with Eqn.1 from different modalities. Then we need to explore the connection $V$ and the hash code $B$. Here we adopt the similar method in [10], and define the following:

$$
B = VR
$$

$$
s.t. R^TR = I, V^TV = nI_r, V1_n = 0_r
$$

where $R \in \mathbb{R}^{n \times n}$ is the connection matrix between $B$ and $V$, and the orthogonal constraint of $R$ makes each column of $R$ independent of each other. And we have enough necessity to embed the label matrix into the learning of hash codes, i.e.,

$$
\min_{B,M} \|B - LM\|_F^2
$$

$$
s.t. B \in \{-1, 1\}^{r \times n}
$$

where $M \in \mathbb{R}^{c \times r}$ is the projection from $L$ to $B$. So we reformulate Eqn.3 in the condition of Eqn.4 and Eqn.5 into the following:

$$
\min_{V,R,M} \|VR(LM)^T - rG^TG\|_F^2 + \|B - LM\|_F^2
$$

$$
s.t. B \in \{-1, 1\}^{r \times n}, R^TR = I
$$
2.1.1. Overall Objective Function

Combining Eqn. 1, and Eqn. 6 together, we have the following objective function of our method:

\[
\min_{V, R, M, B, P} \| VR (LM)^T - rG^T G \|_F^2 + \omega \| B - LM \|_F^2
\]

\[
\sum_{r=0}^l \lambda_c \| \phi (X^{(r)}) - P^{(r)} V \|_F^2
\]

s.t. \( B \in \{-1, 1\}^{r \times n}, R^T R = I, VV^T = nI_r, V1_n = 0_r \) \hspace{1cm} (7)

where \( \omega \) and \( \lambda_c \) are the hyper-parameters.

2.2. Optimization

Directly solving the Eqn. 7 may be very difficult since Eqn. 7 contains the discrete value, and it is an NP problem. So in order to tackle this optimization problem, we proposed an efficient alternative iterative optimization algorithm. At each step, we solve the optimization by only updating one variable while maintaining the others fixed and repeating the procedure until convergence.

1. Optimizing \( P^{(r)} \). When other variables are fixed, Eqn. 7 is then simplified as following formulation:

\[
\min_{P^{(r)}} \| \phi (X^{(r)}) - P^{(r)} V \|_F^2
\]

By taking the partial derivative of Eqn. 7 with respect to \( P^{(r)} \) and setting it to zero, \( P^{(r)} \) can be equivalent to:

\[
P^{(r)} = \phi (X^{(r)}) V^T / n
\]

2. Optimizing \( M \). Here we take the same action as \( P^{(r)} \), we have

\[
M = \left[ (\mu + 1) L^T L \right]^{-1} (rL^T GG^T VR + \mu L^T B)
\]

3. Optimizing \( R \). Given the other variables, the subproblem in Eqn. 7 with respect to \( R \) reduces to the following formulation

\[
\min_R \| VR (LM)^T - rG^T G \|_F^2
\]

s.t. \( R^T R = I \) \hspace{1cm} (11)

Because of Eqn. 11 is an orthogonal rotation problem, so it’s hard to optimize, we develop an alternative version of 11. We directly maximize the similarity matrix and the inner product of asymmetric hashing codes embedding with \( VTV = I_r \), i.e.,

\[
\max_R \text{Tr} (rM^T L^T GG^T R)
\]

s.t. \( R^T R = I \) \hspace{1cm} (12)

Obviously, the current model w.r.t. \( R \) is an OPP problem and can be solved using simple SVD. Let \( P_l \) and \( P_r \) be the left and right singular matrices of \( rMT L^T GG^T \), i.e.,

\[
[P^l, \Sigma, P^r] = \text{svd}(rMT L^T GG^T)
\]

Then we can get the optimal solution of \( R \) is

\[
R = P^l (P^r)^T
\]

4. Optimizing \( V \). To solve the \( V \), we transform it into a matrix trace form with the constraints of \( VV^T = nI_r \) as the following:

\[
\max_R \text{Tr} (rLM^T V^T GG^T + \sum_{r=0}^l \lambda_c P^{(r)} \phi (X^{(r)}) V^T)
\]

s.t. \( VV^T = I \)

To simplify the representation, we define \( J = I_n - \frac{1}{n} 1_n 1_n^T \) and \( Z = rGG^T LM^T R + P^{(r)} \phi (X^{(r)}) \). So with the known constrains, we perform the eigen decomposition of \( ZJZ^T \) like,

\[
ZJZ^T = [Q \bar{Q}] [\Omega \begin{bmatrix} 0 & 0 \end{bmatrix} [Q \bar{Q}]^T
\]

where \( \Omega \in \mathbb{R}^{r \times r} \) and \( Q \in \mathbb{R}^{r \times r} \) are the diagonal matrices of the positive eigenvalues and the corresponding eigenvectors, while \( \bar{Q} \) is the matrix of the remaining \( r - r' \) eigenvectors, and we can easily get the orthogonal matrix \( \bar{Q} \in \mathbb{R}^{r \times (r-r')} \). Then we define \( P = JZ^T \bar{Q} \Omega^{-\frac{1}{2}} \) and a random orthogonal matrix \( \bar{P} \in \mathbb{R}^{n \times (r - r')} \). According to [11], we get the optimal solution of Eqn. 15 as,

\[
V = \sqrt{n} [Q \bar{Q}] [P \bar{P}]^T
\]

5. Optimizing \( B \). Here we take the same action to \( V \). Finally, we have the optimal solution,

\[
B = \text{sgn}(LM)
\]

which the definition of \( \text{sgn}(\cdot) \) is the element-wise indicator operator.

2.3. Hash Functions Learning

Many previous works [12] have shown that two-step hash functions are vital for performance. Namely, we have

\[
\| B - XP_h \|_F^2 + \lambda_h \| P_h \|_F^2
\]

where \( \lambda_h \) is a balance parameter. The optimal \( P \) can be expressed as,

\[
P_h = (X^T X + \lambda_h I)^{-1} X^T B
\]

So the two step ASCMH as,

\[
B_{ts} = \text{sgn}(XP_h)
\]
In this section, we compare our ASCMH method with several state of the art shallow hash methods including CMFH[1], LSSH[16], DCH[17], SRRACH[19], GSPH[20], BATCH[9], SDDH[21] and SLCH[22]. The proposed ASCMH achieves the best performance when $\omega = 0.01$, $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$ on MIRFlickr25K. As in the other paper’s setting, we followed them; all the results averaged over 20 runs.

3.3 Results

In this section, we introduce a brand-new cross-modal hashing technique called Asymmetric Scalable Cross-Modal Hashing (ASCMH). By learning across several modalities, it may investigate consistency and inconsistency. Numerous tests demonstrate ASCMH’s superiority over other SOTA cross-modal retrieval techniques in cross-modal retrieval. In the future, we will expand our approach to deep learning for improved feature extraction and the excellent ability of non-linear fitting.

4. CONCLUSIONS
5. REFERENCES

[1] Guiguang Ding, Yuchen Guo, and Jile Zhou, “Collective matrix factorization hashing for multimodal data,” in CVPR, June 2014.

[2] Tao Yao, Yiru Li, Weili Guan, Gang Wang, Ying Li, Lianshan Yan, and Qi Tian, “Discrete robust matrix factorization hashing for large-scale cross-media retrieval,” IEEE Transactions on Knowledge and Data Engineering, pp. 1–1, 2021.

[3] Dongqing Zhang and Wu-Jun Li, “Large-scale supervised multimodal hashing with semantic correlation maximization,” in AAAI, 2014, vol. 28.

[4] Hua-Junjie Huang, Rui Yang, Chuan-Xiang Li, Yuliang Shi, Shangqin Guo, and Xin-Shun Xu, “Supervised cross-modal hashing without relaxation,” in 2017 IEEE ICME, 2017, pp. 1159–1164.

[5] Shupeng Su, Zhisheng Zhong, and Chao Zhang, “Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval,” in ICCV, 2019, pp. 3027–3035.

[6] Venice Erin Liong, Jiwen Lu, Yap-Peng Tan, and Jie Zhou, “Cross-modal deep variational hashing,” in ICCV, Oct 2017.

[7] Ting Yao, Fuchen Long, Tao Mei, and Yong Rui, “Deep semantic-preserving and ranking-based hashing for image retrieval,” in IJCAI, 2016, vol. 1, p. 4.

[8] Brian Kulis and Trevor Darrell, “Learning to hash with binary reconstructive embeddings,” in NIPS. Citeseer, 2009, vol. 22, pp. 1042–1050.

[9] Yongxin Wang, Xin Luo, Liqiang Nie, Jingkuan Song, Wei Zhang, and Xin-Shun Xu, “Batch: A scalable asymmetric discrete cross-modal hashing,” IEEE Transactions on Knowledge and Data Engineering, 2020.

[10] Zheng Zhang, Zhihui Lai, Zi Huang, Wai Keung Wong, Guo-Sen Xie, Li Liu, and Ling Shao, “Scalable supervised asymmetric hashing with semantic and latent factor embedding,” IEEE Transactions on Image Processing, vol. 28, no. 10, pp. 4803–4818, 2019.

[11] Wei Liu, Cun Mu, Sanjiv Kumar, and Shih-Fu Chang, “Discrete graph hashing,” 2014.

[12] Xin Luo, Liqiang Nie, Xiangnan He, Ye Wu, Zhen-Duo Chen, and Xin-Shun Xu, “Fast scalable supervised hashing,” in The 41st ACM SIGIR, 2018, pp. 735–744.

[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, 2010, pp. 251–260.

[14] 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, 2009, pp. 1–9.

[15] Mark J Huiskes and Michael S Lew, “The mir flickr retrieval evaluation,” in Proceedings of the 1st ACM international conference on Multimedia information retrieval, 2008, pp. 39–43.

[16] 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, New York, NY, USA, 2014, SIGIR ’14, p. 415–424, Association for Computing Machinery.

[17] 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.

[18] Hong Liu, Rongrong Ji, Yongjian Wu, Feiyoue Huang, and Baochang Zhang, “Cross-modality binary code learning via fusion similarity hashing,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7380–7388.

[19] Chuan-Xiang Li, Zhen-Duo Chen, Peng-Fei Zhang, Xin Luo, Liqiang Nie, Wei Zhang, and Xin-Shun Xu, “Scratch: A scalable discrete matrix factorization hashing for cross-modal retrieval,” in Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 1–9.

[20] Devraj Mandal, Kunal N Chaudhury, and Soma Biswas, “Generalized semantic preserving hashing for n-label cross-modal retrieval,” in CVPR, 2017, pp. 4076–4084.

[21] Jianyang Qin, Lunke Fei, Jian Zhu, Jie Wen, Chunwei Tian, and Shuai Wu, “Scalable discriminative discrete hashing for large-scale cross-modal retrieval,” in ICASSP, IEEE, 2021, pp. 4330–4334.

[22] Heng Tao Shen, Luchen Liu, Yang Yang, Xing Xu, Zi Huang, Fumin Shen, and Richang Hong, “Exploiting subspace relation in semantic labels for cross-modal hashing,” IEEE Transactions on Knowledge and Data Engineering, 2020.