Joint Dictionary Learning and Semantic Constrained Latent Subspace Projection for Cross-Modal Retrieval

Jianlong Wu, Zhouchen Lin✉, Hongbin Zha
Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, Beijing 100871, P. R. China
Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai 200240, P. R. China
{jlwu1992,zlin}@pku.edu.cn,zha@cis.pku.edu.cn

ABSTRACT

With the increasing of multi-modal data on the internet, cross-modal retrieval has received a lot of attention in recent years. It aims to use one type of data as query and retrieve results of another type. For different modality data, how to reduce their heterogeneous property and preserve their local relationship are two main challenges. In this paper, we present a novel joint dictionary learning and semantic constrained latent subspace learning method for cross-modal retrieval (JDSLC) to deal with above two issues. In this unified framework, samples from different modalities are encoded by their corresponding dictionaries to reduce the semantic gap. In the meantime, we learn modality-specific projection matrices to map the sparse coefficients into the shared latent subspace. Meanwhile, we impose a novel cross-modal similarity constraint to make the representations of samples that belong to same class but from different modalities as close as possible in the latent subspace. An efficient algorithm is proposed to jointly optimize the proposed model and learn the optimal dictionary, coefficients and projection matrix for each modality. Extensive experimental results on multiple benchmark datasets show that our proposed method outperforms the state-of-the-art approaches.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; *Structure and multilingual text search*;

KEYWORDS

Cross-modal retrieval; Semantic constraints; Dictionary learning

ACM Reference Format:
Jianlong Wu, Zhouchen Lin✉, Hongbin Zha. 2018. Joint Dictionary Learning and Semantic Constrained Latent Subspace Projection for Cross-Modal Retrieval . In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3269206.3269296

1 INTRODUCTION

Information retrieval is an important task in computer science. Single modality retrieval has been well studied, such as image retrieval, text retrieval and etc. Recently, with the rapid increase of multimedia data which consists of multi-modality information, people pay much attention to cross-modal retrieval, which enables us to use one type data as query and retrieve relevant samples from the database formed by data of other modalities. Towards this task, many works have been proposed recently. For a comprehensive review, please refer to [13]. For data from different modalities, they share the same underlying content, but there are also semantic gaps and heterogeneous properties. It is very challenging to measure their cross-modal similarity directly.

For cross-modal retrieval, there are mainly three issues we should take into consideration. First of all, we need to reduce the semantic gap between different type data. To tackle this issue, dictionary learning methods achieved very good performance. Secondly, it is crucial to measure their similarity and distance in a common space. Latent subspace learning methods are the most popular approaches towards this issue. It can efficiently compute their similarity by mapping various modality data into one shared latent subspace. Thirdly, in the shared space, distance between samples of same category should be as close as possible. However, according to [3], common space projection cannot promise this. It is necessary to add a similarity constraint to preserve the local relationship after projecting into the latent subspace.

All these properties are very necessary and important for cross-modal retrieval. However, existing cross-modal retrieval methods mainly focus on the second part of the above issues and adopt different constraints to achieve the desired properties. In this paper, we propose a novel unified framework to simultaneously learn the dictionary for coding and matrices for projecting to the shared space with a novel semantic cross-modal similarity constraint. On the one hand, ℓ_1-norm regularized dictionary learning (also known as LASSO) is adopted to learn sparse codes for each type data and reduce the heterogeneous property at the primary stage. On the other hand, we map different modality data into a shared latent space, which is learnt by graph embedding instead of using the simple label space. So that we can conveniently measure their similarity in this common space. Please remind that we hope the distance between samples of same category should be as close as possible, but the common space projection cannot promise this according to [3]. So we exert a novel cross-modal similarity constraint on the representations of cross-modal data in the shared latent space. Finally, we combine the above terms together to learn the optimal dictionary and the projection matrix.

The proposed JDSLC has three main contributions. Firstly, we propose a novel unified framework to simultaneously learn related variables. Secondly, a novel cross-modal constraint is proposed to well preserve the relationship among samples of same class.

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3269296

Finally, we propose an iterative algorithm to efficiently solve the proposed problem. Experimental results on related datasets show the superiority of our proposed method.

2 RELATED WORK

For cross-modal retrieval, extensive research has been proposed and many methods achieve the state-of-the-art performance. We briefly review some relevant methods, including latent subspace learning [11] and dictionary learning [4, 15].

Latent subspace learning is the most popular method for cross-modal retrieval. By projecting multi-modal data into one common subspace, we can measure their similarity efficiently. Canonical Correlation Analysis (CCA) [2], Partial Least Squares (PLS) and Bilinear Model (BLM) are three main classic unsupervised methods. Despite the above unsupervised classic methods, there are also many approaches that incorporate label information to facilitate the retrieval, where the latent subspace is often defined by their semantic label. Sharma et al. [10] presented a general multi-view feature extraction approach called generalized multiview analysis (GMA) which extended linear discriminant analysis and marginal Fisher analysis (MFA) to their multiview cases. Wang et al. [12] proposed a method to learn coupled feature space with ℓ_{21}-norm projection matrix penalty and low-rank constraint on the projected data. Then they employed a multi-modal graph regularization term to preserve the local relationship [11]. In [3, 16], a joint representation learning method was presented to explore the influence of pairwise constraint during latent space regression. Kang et al. [6] added a local group-based priori and a $\epsilon-$dragging term for robust representation. Instead of using the simple label information as the latent subspace, Wu et al. [14] came up with a joint latent subspace learning and regression method to learn the optimal common subspace for projection.

Dictionary learning is often adopted for this task. Huang et al. [4] proposed a coupled dictionary and feature space learning method, which learned a pair of dictionaries for describing cross-domain image data and explored the correlation between sparse codes of different modality. Inspired by [5] and based on the discriminative dictionary learning method, Deng et al. [1] came up with an approach that adopted a common label alignment within the class label space to augment the correlations among all the modalities.

Besides these two kinds of methods, there are some other methods, such as deep learning methods, rank based methods, and etc.

However, most existing methods only focus on part of the necessary issues for cross-modal retrieval. It is very necessary to come up with a unified framework that takes all these important factors into consideration. So we propose a joint dictionary learning and latent subspace learning framework with a novel cross-modal similarity constraint to reduce the semantic gap, measure their similarity efficiently, and preserve the local relationship.

3 PROPOSED APPROACH

3.1 Problem Formulation

For the task of cross-modal retrieval, we mainly consider the situation with two different modalities, such as the most popular image versus text retrieval. Denote $X^a = [x_1^a, \cdots, x_N^a] \in \mathbb{R}^{d_a \times N}$ and $X^b = [x_1^b, \cdots, x_N^b] \in \mathbb{R}^{d_b \times N}$ as the N feature pairs extracted from two different domains. On the one hand, we hope to adopt sparse

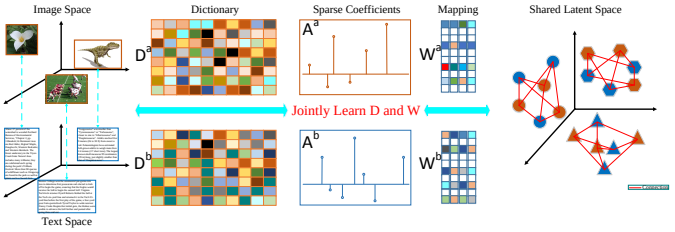

Figure 1: Framework of the proposed JDSLC method.

representation method to learn the dictionary $D^I \in \mathbb{R}^{d_I \times K}$ ($I \in \{a, b\}$) with K atoms and sparse coefficient $A^I \in \mathbb{R}^{K \times N}$ for features of each domain. On the other hand, we expect to learn projection matrices $W^I \in \mathbb{R}^{K \times C}$ to map the sparse representations of different modality into one common latent space Y to better measure their cross-modal similarity. Instead of directly using label matrix to define Y, we learn the orthogonal space Y by spectral regression [14]:

$$\min_{Y} \frac{1}{2} \sum_{i,j} \|y_i - y_j\|_2^2 S_{ij} = tr\left(Y^T(H - S)Y\right) \quad s.t.\ Y^T Y = \mathbf{I}. \quad (1)$$

Here, weight S_{ij} is equal to 1 only when the i-th sample has the same class information with the j-th sample, and otherwise $S_{ij} = 0$. H is a diagonal matrix with the i-th diagonal element as $H_{ii} = \sum_{j=1}^{N} S_{ij}$. The above problem can be simply solved by eigenvalue decomposition. However, separating the dictionary learning from the common space projection might make both the dictionary D and the projection direction W suboptimal. In this case, in our proposed model, we jointly learn the discriminative dictionary and projection direction. According to [3], common subspace mapping cannot guarantee the distances among samples of same class are small. So it is necessary to incorporate a cross-modal similarity term to constrain the representation of different modality data in the latent subspace. Then the objective function for our proposed method can be formulated as follows:

$$\min_{D,A,W} \sum_{I \in \{a,b\}} \left(\left\|X^I - D^I A^I\right\|_F^2 + \alpha \left\|A^I\right\|_1 + \beta \left\|Y - (W^I)^T A^I\right\|_F^2 + \gamma \left\|W^I\right\|_{2,1} \right) + \lambda \Omega\left((W^a)^T A^a, (W^b)^T A^b\right), \quad (2)$$

where α, β, γ, and λ are balance parameters to control the relative contribution of each item. To enforce sparsity, ℓ_1 norm is used to constrain the sparse code A. For a matrix U, the $\ell_{2,1}$-norm is defined as the sum of the ℓ_2-norm of the rows of U: $\|U\|_{2,1} = \sum_{i=1}^{m} \|U^{(i)}\|_2$. We apply $\ell_{2,1}$ norm to the projection matrix W for feature selection. We present the pipeline of our framework in Figure 1.

In the objective function of Eq. (2), the first term minimizes reconstruction error of dictionary learning, and the third term minimizes projection error, while the final term minimizes the distance between representations of different domains in shared latent space.

For the cross-modal similarity constraint Ω, the commonly used form is the pairwise constraint [3]:

$$\Omega\left((W^a)^T A^a, (W^b)^T A^b\right) = \left\|(W^a)^T A^a - (W^b)^T A^b\right\|_F^2. \quad (3)$$

However, this kind of constraint can only promise representations of pairwise samples in the latent space could be close. Please remind that for cross-modal retrieval, we hope representations of not only

pairwise samples, but also samples belong to the same class should be as close as possible in the common subspace. So we further propose a novel cross-modal similarity constraint as follows:

$$\Omega\left((W^a)^T A^a,(W^b)^T A^b\right)=\sum_{i=1}^{N}\sum_{j=1}^{N} S_{ij}\left\|(W^a)^T A_i^a-(W^b)^T A_j^b\right\|_2^2. \quad (4)$$

Here, S is same to that in Eq. (1).

Then the final objective function for our JDSLC model is:

$$\min_{D,A,W}\sum_{I\in\{a,b\}}\left(\left\|X^I-D^I A^I\right\|_F^2+\alpha\left\|A^I\right\|1+\beta\left\|Y-(W^I)^T A^I\right\|_F^2 +\gamma\left\|W^I\right\|_{2,1}\right)+\lambda\sum_{i=1}^{N}\sum_{j=1}^{N} S_{ij}\left\|(W^a)^T A_i^a-(W^b)^T A_j^b\right\|_2^2. \quad (5)$$

3.2 Optimization

It's obvious that the optimization problem in Eq. (5) is not jointly convex to D, A, and W. So it is difficult to optimize jointly. However, it is convex to each variable while other variables are fixed. In the following, we present an iterative algorithm to optimize the dictionaries D, sparse codes A, and projection directions W, respectively.

We first update D by fixing A and W as constants. The problem of optimizing D can be formulated as

$$\min_{D^I}\left\|X^I-D^I A^I\right\|_F^2,\quad I\in\{a,b\}.\ s.t.\quad \left\|d_i^I\right\|_2\le 1,\forall i, \quad (6)$$

which is a quadratically constrained quadratic program (QCQP) problem with respect to D^I. It can be solved by the Lagrange dual techniques [7].

Then, with the dictionaries D and projection matrices W fixed, we calculate the sparse codes A. The problem in Eq. (5) is transformed into the following problem:

$$\min_{A^I}\left\|X^I-D^I A^I\right\|_F^2+\alpha\left\|A^I\right\|_1+\beta\left\|Y-(W^I)^T A^I\right\|_F^2 +\lambda\sum_{i=1}^{N}\sum_{j=1}^{N} S_{ij}\left\|(W^a)^T A_i^a-(W^b)^T A_j^b\right\|_2^2. \quad (7)$$

For the last term in Eq. (7), we can expand and simplify it as:

$$\begin{aligned}
&\arg\min_{A^a}\lambda\sum_{i=1}^{N}\sum_{j=1}^{N} S_{ij}\left\|(W^a)^T A_i^a-(W^b)^T A_j^b\right\|_2^2\\
&=\arg\min_{A^a}\lambda\sum_{i=1}^{N}\left(\left(\sum_{j=1}^{N} S_{ij}\right)(W^a)^T A_i^a(A_i^a)^T W^a-\right.\\
&\left.(W^a)^T A_i^a\sum_{j=1}^{N}\left(S_{ij}(A_j^b)^T W^b\right)-\sum_{j=1}^{N}\left(S_{ij}(W^b)^T A_j^b\right)(A_i^a)^T W^a\right)\\
&=\arg\min_{A^a}\lambda\sum_{i=1}^{N}\left\|\frac{\sum_{j=1}^{N}\left(S_{ij}(W^b)^T A_j^b\right)}{\sqrt{\sum_{j=1}^{N} S_{ij}}}-\sqrt{\sum_{j=1}^{N} S_{ij}}(W^a)^T A_i^a\right\|_2^2\\
&=\arg\min_{A^a}\lambda\sum_{i=1}^{N}\left\|Z_i^b-\sqrt{\sum_{j=1}^{N} S_{ij}}(W^a)^T A_i^a\right\|_2^2,
\end{aligned} \quad (8)$$

where $Z_i^b=\frac{\sum_{j=1}^{N}\left(S_{ij}(W^b)^T A_j^b\right)}{\sqrt{\sum_{j=1}^{N} S_{ij}}}$. Similarly, with $Z_i^a=\frac{\sum_{j=1}^{N}\left(S_{ij}(W^a)^T A_j^a\right)}{\sqrt{\sum_{j=1}^{N} S_{ij}}}$, the last term in Eq. (7) during optimizing A^b is equal to:

$$\begin{aligned}
&\min_{A^b}\lambda\sum_{i=1}^{N}\sum_{j=1}^{N} S_{ij}\left\|(W^a)^T A_i^a-(W^b)^T A_j^b\right\|_2^2\\
&=\min_{A^b}\lambda\sum_{i=1}^{N}\left\|Z_i^a-\sqrt{\sum_{j=1}^{N} S_{ij}}(W^b)^T A_i^b\right\|_2^2.
\end{aligned} \quad (9)$$

To compute A^a and A^b, by combining the first, third, and fourth terms in Eq. (7) into one term, the problem can be simplified into:

$$\begin{aligned}
&\min_{A_i^a}\left\|\begin{pmatrix} X_i^a\\ \sqrt{\beta}Y_i\\ \sqrt{\lambda}Z_i^b\end{pmatrix}-\begin{pmatrix} D^a\\ \sqrt{\beta}(W^a)^T\\ \sqrt{\lambda\sum_{j=1}^{N} S_{ij}}(W^a)^T\end{pmatrix}A_i^a\right\|_2^2+\alpha\left\|A_i^a\right\|_1,\\
&\min_{A_i^b}\left\|\begin{pmatrix} X_i^b\\ \sqrt{\beta}Y_i\\ \sqrt{\lambda}Z_i^a\end{pmatrix}-\begin{pmatrix} D^b\\ \sqrt{\beta}(W^b)^T\\ \sqrt{\lambda\sum_{j=1}^{N} S_{ij}}(W^b)^T\end{pmatrix}A_i^b\right\|_2^2+\alpha\left\|A_i^b\right\|_1,
\end{aligned} \quad (10)$$

where $i\in\{1,2,\cdots,N\}$. It is a standard ℓ_1 norm regularized sparse coding problem. For sparse coefficients A_i^I of each sample, we can use the SPAMS toolbox [8] to derive solutions.

Finally, we need to update the projection directions W by fixing D and A. Take W^a for example, we have:

$$\begin{aligned}
&\min_{W^a}J(W^a)=\beta\left\|Y^T-(A^a)^T W^a\right\|_F^2+\gamma\left\|W^a\right\|_{2,1}\\
&+\lambda\left(tr\left((W^a)^T A^a H(A^a)^T W^a\right)-tr\left((W^a)^T A^a S(A^b)^T W^b\right)\right.\\
&\left.-tr\left((W^b)^T A^b S(A^a)^T W^a\right)\right).
\end{aligned} \quad (11)$$

H is also same to that in Eq. (1). Then the derivative of $J(W^a)$ with respect to W^a is:

$$\begin{aligned}
\frac{\partial J(W^a)}{\partial W^a}=&-2\beta A^a\left(Y^T-(A^a)^T W^a\right)+2\gamma Q^a W^a\\
&+2\lambda A^a\left(H(A^a)^T W^a-S(A^b)^T W^b\right),
\end{aligned} \quad (12)$$

where Q^a is a diagonal matrix with the i-th diagonal element as $Q_{ii}^a=\frac{1}{\sqrt{\|w_i^a\|_2^2+\epsilon}}$. Here, ϵ is a very small constant to avoid the denominator being 0. By setting the above derivative in Eq. (12) to zero, we can get:

$$W^a=\left(\beta A^a(A^a)^T+\lambda A^a H(A^a)^T+\gamma Q^a\right)^{-1}\left(\beta A^a Y^T+\lambda A^a S(A^b)^T W^b\right).$$

Similarly, we can compute W^b by:

$$W^b=\left(\beta A^b(A^b)^T+\lambda A^b H(A^b)^T+\gamma Q^b\right)^{-1}\left(\beta A^b Y^T+\lambda A^b S(A^a)^T W^a\right).$$

We repeat the above three steps to alternatively optimize D, A, and W until the objective value of Eq. (5) converges, when the rate of change between two iterations is less than a small threshold.

After learning the optimal variables based on the training samples, we map the testing samples of different modalities into the common subspace with the modality-specific projection matrices. So that we can measure their similarity and retrieve the related cross-modal samples.

Table 1: MAP Comparison on the Wikipedia dataset.

Methods	Image query	Text query	Average
CCA [2]	0.2549	0.1846	0.2198
GMMFA [10]	0.2750	0.2139	0.2445
GMLDA [10]	0.2751	0.2098	0.2425
PL-Ranking [17]	0.2625	0.2221	0.2423
LCFS [12]	0.2798	0.2141	0.2470
DDLCC [1]	0.2909	0.2261	0.2585
JFSSL [11]	0.3063	0.2275	0.2669
LGCFL [6]	0.3009	0.2377	0.2693
JLSLR [14]	0.3168	0.2346	0.2757
JDSLC	**0.3177**	**0.2531**	**0.2854**

Table 2: MAP Comparison on the MIR-Flickr dataset.

Methods	Image query	Text query	Average
CCA [2]	0.1455	0.1438	0.1447
GMMFA [10]	0.2657	0.1884	0.2271
GMLDA [10]	0.2662	0.1893	0.2278
PL-Ranking [17]	0.2851	0.2323	0.2587
LCFS [12]	0.3860	0.2658	0.3259
DDLCC [1]	0.3925	0.2861	0.3393
JFSSL [11]	0.4122	0.2802	0.3462
LGCFL [6]	0.4060	0.2816	0.3438
JLSLR [14]	0.4131	0.2831	0.3481
JDSLC	**0.4179**	**0.2904**	**0.3541**

4 EXPERIMENTS AND RESULTS

We test the performance of our JDSLC method on two datasets, including the Wikipedia dataset and the MIR-Flickr dataset.

The Wikipedia dataset contains 2, 866 image-text pairs, which are generated from the featured article of Wikipedia. There are 10 semantic categories in total. For each pair, the text is a long article describing the label related information, and the image is high correlated to the content of the article. We adopt the same setting as that in [11, 12], which splits 2, 866 pairs into a training set of 1, 300 pairs (130 pairs per class) and a testing set of 1, 566 pairs. For text features, latent Dirichlet allocation (LDA) is used to extract 10 dimensions representation. For image representation, we extract the 128 dimensional SIFT descriptor histograms.

The MIR-Flickr dataset contains 25, 000 image-tag pairs. We select image-tags pairs that exclusively belong to only one of the 10 largest concepts, which results in 5, 730 pairs in total for our experiments [17]. We directly adopt the 500-dimensional bag of words feature vectors based on SIFT descriptions and 1000-dimensional word frequency feature vectors to represent images and textual tags, respectively. We adopt same setting as that in [17]. 75% of the data are selected as training samples, and the remaining for testing.

We adopt the commonly used mean average precision (MAP) to evaluate the performance. For details of the MAP computation, please refer to [9]. Higher MAP scores show better result.

For parameters α, β, γ, and λ of our proposed JDSLC method in Eq. (5), we fine tune them by searching the gird of $\{10^{-2}, 10^{-1}, \cdots, 10^{3}\}$ based on cross validation.

In Tables 1 and 2, we show the results of these start-of-the-art methods on these two datasets. We can see that our JDSLC method achieves the best performance on both two datasets. On Wikipedia dataset, the average MAP of our method is 0.2854, while the second best result is 0.2757. There is 3.5% improvement relatively on this dataset. On MIR-Flickr dataset, we achieve 0.3541, which is 1.7% higher than the second best result achieved by the JLSLR [14]. We also test the significance between the proposed JDSLC and JLSLR [14], which achieves the second best results. The p-value on these two datasets between these two methods are 0.03 and $5.1*10^{-4}$, respectively. Both of them are less than 0.05, which shows our result has significant difference with that of JLSLR. We can also observe the similar results when compared with other methods. Based on the results, we can see that the dictionary learning and local semantic constraint work very well for this task. With joint heterogeneous property reducing and local relationship preserving, our method achieves the best performance.

5 CONCLUSION

For cross-modal retrieval, we propose a novel joint dictionary learning and latent subspace projection method with local semantic constraint. Experimental results show the superiority of our method.

6 ACKNOWLEDGEMENTS

Zhouchen Lin is supported by National Basic Research Program of China (973 Program) (Grant no. 2015CB352502), National Natural Science Foundation (NSF) of China (Grant nos. 61625301 and 61731018), Qualcomm, and Microsoft Research Asia. Hongbin Zha is supported by Beijing Municipal Natural Science Foundation (Grant no. 4152006).

REFERENCES

[1] C. Deng, X. Tang, J. Yan, W. Liu, and X. Gao. Discriminative dictionary learning with common label alignment for cross-modal retrieval. *IEEE TMM*, 18(2):208–218, 2016.
[2] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.
[3] R. He, M. Zhang, L. Wang, Y. Ji, and Q. Yin. Cross-modal subspace learning via pairwise constraints. *IEEE TIP*, 24(12):5543–5556, 2015.
[4] D.-A. Huang and Y.-C. Frank Wang. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *ICCV*, pages 2496–2503, 2013.
[5] Z. Jiang, Z. Lin, and L. S. Davis. Label consistent k-svd: Learning a discriminative dictionary for recognition. *IEEE TPAMI*, 35(11):2651–2664, 2013.
[6] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan. Learning consistent feature representation for cross-modal multimedia retrieval. *IEEE TMM*, 17(3):370–381, 2015.
[7] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *NIPS*, 2007.
[8] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *ICML*, 2009.
[9] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM Multimedia*, 2010.
[10] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, 2012.
[11] K. Wang, R. He, L. Wang, W. Wang, and T. Tan. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE TPAMI*, 38(10):2010–2023, 2016.
[12] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. Learning coupled feature spaces for cross-modal matching. In *ICCV*, 2013.
[13] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.
[14] J. Wu, Z. Lin, and H. Zha. Joint latent subspace learning and regression for cross-modal retrieval. In *ACM SIGIR*, 2017.
[15] X. Xu, Y. Yang, A. Shimada, R.-i. Taniguchi, and L. He. Semi-supervised coupled dictionary learning for cross-modal retrieval in internet images and texts. In *ACM Multimedia*, pages 847–850, 2015.
[16] X. Zhai, Y. Peng, and J. Xiao. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(6):965–978, 2014.
[17] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian. Pl-ranking: a novel ranking method for cross-modal retrieval. In *ACM Multimedia*, pages 1355–1364, 2016.