Cross-modal Retrieval with Label Completion

Xing Xu
University of Electronic Science and Technology of China
Chengdu, China
xing.xu@uestc.edu.cn

Fumin Shen
University of Electronic Science and Technology of China
Chengdu, China
fumin.shen@gmail.com

Yang Yang
University of Electronic Science and Technology of China
Chengdu, China
dlyyang@gmail.com

Heng Tao Shen
The University of Queensland
Brisbane, Australia
University of Electronic Science and Technology of China
Chengdu, China
shenhengtao@hotmail.com

Li He
Qualcomm R&D Center
San Diego, USA
lih@qti.qualcomm.com

Jingkuan Song
University of Trento
Trento, Italy
jingkuan.song@unitn.it

ABSTRACT

Cross-modal retrieval has been attracting increasing attention because of the explosion of multi-modal data, e.g., texts and images. Most supervised cross-modal retrieval methods learn discriminant common subspaces minimizing the heterogeneity of different modalities by exploiting the label information. However, these methods neglect the fact that, in practice, the given labels of training data might be incomplete (i.e., some of their labels are missing). The low-quality labels result in less effective subspace and consequent unsatisfactory retrieval performance. To tackle this, we propose a novel model that simultaneously performs label completion and cross-modal retrieval. Specifically, we assume the to-be-learned common subspace can be jointly derived through two aspects: 1) linear projection from modality-specific features and 2) enriching mapping from the incomplete labels. We thus formulate the subspace learning problem as a co-regularized learning framework based on multi-modal features and incomplete labels. Extensive experiments on two large-scale multi-modal datasets demonstrate the superiority of our model for both label completion and cross-modal retrieval over the state-of-the-arts.

Keywords

Cross-modal retrieval; label completion

1. INTRODUCTION

The past decade has witnessed the explosion of online imagery contents, especially on social photo-sharing websites such as Facebook, Flickr and Instagram. Usually, large-scale Internet photo collections consist of multi-modal data of images and texts. As the examples shown in Fig. 1, to describe the content of "Image", two types of textual data are usually associated: 1) "Text" that refers to descriptions from the surrounding web pages of the image or user-provided coarse tags; and 2) "Label" that represents the high-level semantic labels manually annotated by human annotators. Earlier research works, such as image search [30, 31, 21, 22] or document retrieval [5], try to explore the distinct characteristics in individual modality of "Image" or "Text". Recently, jointly modeling the statistics of images and associated textual data has continuously attracted much attention. Typical applications, such as automatic image annotation/caption [1, 17, 29, 27] and keyword-based image search [8, 23, 16], aim to build direct connection from "Image" to "Label" or the opposite. Due to the distinct statistical properties of "Image" and "Text", there have been increasing interests and efforts to model bidirectional connections across these two modalities. In this paper, we consider cross-modal retrieval problem on large-scale Image-Text datasets, where queries from one modality (e.g., "Image") are matched to database entries from another (e.g., "Text"). Since "Image" and "Text" reside in different feature spaces, the core issue of cross-modal retrieval is how to eliminate the diversity between the heterogeneous features.

To this end, many approaches have been proposed to learn a common latent subspace for cross-modal retrieval, where the projected features of different modalities are homogeneous and can be directly matched. In general, these methods can be classified into two categories: unsupervised and supervised. The unsupervised methods, including the classical ones such as Canonical Correlation Analysis (CCA) [12], Partial Least Square (PLS) [20] and their extensions [7, 9, 13, 19, 25], aim to directly build the correspondence and preserve the correlation of Image-Text pairs in the learned subspace. However, these approaches ignore the valuable label information ("Label") associated with the Image-Text pairs, resulting in less discriminative subspace.

On the other hand, the supervised methods [14, 24, 28] learn discriminant subspace from different modalities by fur-

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2967231

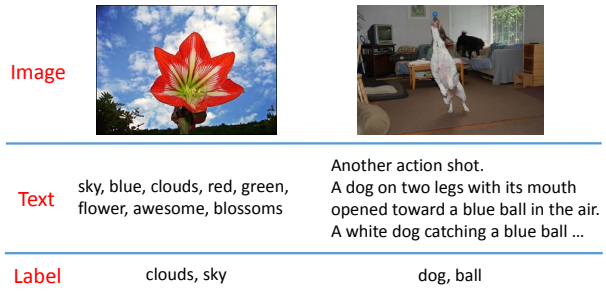

Figure 1: Two exemplars in Image-Text datasets: NUS-WIDE [3] (left) and Flickr30K [32] (right).

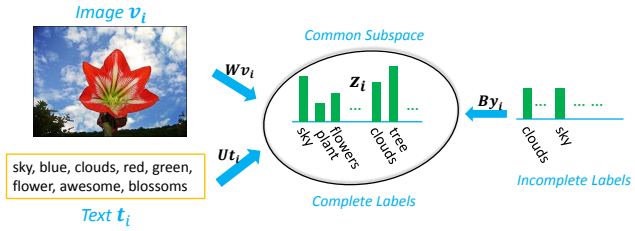

Figure 2: The joint learning framework of the proposed method.

ther exploiting the cues of "Label". However, most of these methods assume that each Image-Text pair contains a single label, which is not usually satisfied in practical Image-Text data with multi-label assumption as shown in Fig. 1. To overcome the shortcoming of these single-label approaches, several recent studies [4, 28] restrict the subspace to be decided by the label information and learn it under a linear classification framework with multi-label assumption. The latest work of [18] further extends the classic CCA to the multi-label situation under the supervision of multi-label information, achieving improved performance for cross-modal retrieval.

However, in practice the given labels are usually incomplete and insufficient as semantic description for corresponding images due to negligence or mistakes of the human annotators. For example, in Fig. 1 other proper labels such as $\langle flowers, tree, plant\rangle$, $\langle action, playing, room\rangle$, are missing for the two images respectively. Directly using these incomplete labels for subspace learning may not guarantee to achieve effective subspace for cross-modal retrieval. Therefore, it is necessary to first obtain complete labels via label completion before learning the discriminative subspace. Indeed, label completion is an attractive research topic in automatic image annotation problem. However, the existing studies [2, 15, 26] of label completion mainly focus on identifying the correct associations between the unimodal "Image" and "Label", not really suitable for multi-modal data containing both "Image" and "Text".

To tackle these challenges, we propose a novel method (as shown in Fig. 2) that performs label completion and cross-modal retrieval simultaneously. For cross-modal retrieval, linear regression is used to project data from different modalities into a common subspace; at the same time, for label completion, the common subspace is assumed to be defined exactly by the ideally complete label information, into which the incomplete labels can be transformed via enriching mapping. The two processes are integrated into a joint learning framework ensuring that the learned common subspace not only captures the intra- and inter-modality discrimination but also accounts for the ideally complete label information.

The main contributions of our work can be summarized as follows: 1) We propose a novel method to simultaneously tackle label completion and cross-modal retrieval problems. 2) Our method utilizes multi-modal data of image and text for label completion, which is an novel extension of traditional label completion approaches based on unimodal image data. 3) An iterative algorithm is presented to efficiently solve the complex minimization problem and can be applied to large-scale training data. 4) Experimental results on two large-scale multi-modal datasets have shown that our method obtains promising results on both label completion and cross-modal retrieval task.

2. THE PROPOSED METHOD

2.1 Problem Formulation

Assume that the multi-modal training data consists of n instances with image-text pairs, i.e. $\mathcal{X} = \{x_i\}_{i=1}^n$, $x_i = (\mathbf{v}_i, \mathbf{t}_i)$, where $\mathbf{v}_i \in \mathbb{R}^m$ is the i-th column of image feature matrix $\mathbf{V} \in \mathbb{R}^{m\times n}$, and $\mathbf{t}_i \in \mathbb{R}^d$ is the i-th column of the text feature matrix $\mathbf{T} \in \mathbb{R}^{d\times n}$. Here m and d are the dimensionality of image and text feature space, respectively. In addition, the incomplete labels $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n] \in \mathbb{R}^{c\times n}$ of training instances $\mathcal{X}$ are also given, where c is the total number of labels and $\mathbf{y}_i = [y_{i,1}, ..., y_{i,c}] \in \mathbb{R}^c$ is the label vector for the i-th instance x_i. Note that, if x_i has the j-th label, then $y_{i,j} = 1$, otherwise $y_{i,j} = 0$. And x_i may have a single label or multiple labels. Unlike the previous subspace learning approaches [4, 18, 28] that define the common subspace by the given incomplete label information, here we consider to learn a common subspace that is defined by the ideally complete label information. The features of different modalities and the given incomplete labels can be simultaneously mapped to the subspace.

Specifically, our goal is to obtain a real-valued matrix $\mathbf{Z} \in \mathbb{R}^{c\times n}$ that satisfies the following three conditions simultaneously: 1) the i-th column vector $\mathbf{z}_i$ in $\mathbf{Z}$ is the common subspace feature representation of instance x_i, and it eliminates the heterogeneity between different features of $\mathbf{v}_i$ and $\mathbf{t}_i$; 2) $\mathbf{z}_i$ is sufficiently consistent with the provided incomplete labels $\mathbf{y}_i$, i.e. when performing label completion based on $\mathbf{z}_i$, $\mathbf{y}_i$ should be a subset of the ideally complete labels; 3) $\mathbf{z}_i$ preserves the discriminative properties of the ideally label information of different classes. We incorporate the three criteria simultaneously and integrate the label completion and cross-modal retrieval into a joint learning framework.

In particular, our learning framework (as shown in Fig. 2) contains two processes: 1) training modality-specific projections $\mathbf{W}\mathbf{v}_i \to \mathbf{z}_i$ and $\mathbf{U}\mathbf{t}_i \to \mathbf{z}_i$ to obtain the common subspace representation $\mathbf{z}_i$ from $\mathbf{v}_i$ and $\mathbf{t}_i$ of image and text modalities, respectively; 2) training an enriching mapping $\mathbf{B}\mathbf{y}_i \to \mathbf{z}_i$ to complete the provided incomplete labels $\mathbf{y}_i$ by recalling the missing labels that are likely to co-occur with those existing in $\mathbf{y}_i$. Therefore, for all the n instances, the loss function of the two sub-tasks can be written as

$$\frac{1}{n}\sum_{i=1}^{n}(\|\mathbf{z}_i - \mathbf{W}\mathbf{v}_i\|^2 + \|\mathbf{z}_i - \mathbf{U}\mathbf{t}_i\|^2 + \|\mathbf{z}_i - \mathbf{B}\mathbf{y}_i\|^2). \quad (1)$$

Here $\mathbf{W} \in \mathbb{R}^{c\times d}$ and $\mathbf{U} \in \mathbb{R}^{c\times m}$ are the projection matrices of image and text features, respectively, and $\mathbf{B} \in \mathbb{R}^{c\times c}$ is the enriching mapping matrix of given incomplete labels. To simplify Eq. 1, for each x_i, we force the output of the two sub-tasks to agree on $\mathbf{z}_i$, resulting in a *cross-modal co-regularized learning* problem by minimizing:

$$\min_{W,V,B} \frac{1}{n}\sum_{i=1}^{n}(\|\mathbf{B}\mathbf{y}_i - \mathbf{W}\mathbf{v}_i\|^2 + \|\mathbf{B}\mathbf{y}_i - \mathbf{U}\mathbf{t}_i\|^2). \quad (2)$$

In fact, the problem of Eq. 2 has a trivial solution at $\mathbf{B} = \mathbf{W} = \mathbf{U} = \mathbf{0}$. Therefore, it is necessary to add regularization terms for the model parameters $\mathbf{W}$, $\mathbf{U}$, and $\mathbf{B}$. In particular, for $\mathbf{W}$ and $\mathbf{U}$, we can simply add ℓ_2 regularizer for them due to their linear regression forms; for $\mathbf{B}$, because it is originally introduced to ensure the consistence between the given incomplete labels and the ideally complete labels during enriching mapping, it may be under constrained due to the absence of the ideally complete labels.

Inspired by [2], we constrain $\mathbf{B}$ by expanding the condition of enriching mapping, i.e. for the given incomplete labels $\mathbf{y}_i$ of instance x_i, $\mathbf{B}$ not only enriches $\mathbf{y}_i$ to its ideally complete labels, but also enriches a "corrupted" version of $\mathbf{y}_i$ (denoted as $\hat{\mathbf{y}}_i$) to $\mathbf{y}_i$. The idea is that if the consistence between $\mathbf{y}_i$ and its "corrupted" labels $\hat{\mathbf{y}}_i$ matches the consistence between $\mathbf{y}_i$ and its ideally complete labels, then applying $\mathbf{B}$ to $\mathbf{y}_i$ would recover the ideally complete labels.

Suppose that the "corrupted" labels $\hat{\mathbf{y}}_i$ are created by randomly removing each label in $\mathbf{y}_i$ with probability $p > 0$, the expected error of the enriching mapping between $\hat{\mathbf{y}}_i$ and $\mathbf{y}_i$ can be expressed as $\mathbb{E}[\|\mathbf{y}_i - \mathbf{B}\hat{\mathbf{y}}_i\|^2]_{p(\hat{\mathbf{y}}_i|\mathbf{y}_i)}$ under the corrupting distribution of $p(\hat{\mathbf{y}}_i|\mathbf{y}_i)$. Then the total expected error of all n instances can be computed as:

$$r(\mathbf{B}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\mathbf{y}_i - \mathbf{B}\hat{\mathbf{y}}_i\|^2]_{p(\hat{\mathbf{y}}_i|\mathbf{y}_i)}. \quad (3)$$

By defining $\mathbf{P} = \sum_{i=1}^{n}\mathbf{y}_i\mathbb{E}[\hat{\mathbf{y}}_i]^\top$ and $\mathbf{Q} = \sum_{i=1}^{n}\mathbb{E}[\mathbf{y}_i\mathbf{y}_i^\top]$, Eq. 3 can be expanded into

$$r(\mathbf{B}) = \frac{1}{n}trace(\mathbf{B}\mathbf{Q}\mathbf{B}^\top - 2\mathbf{P}\mathbf{B}^\top + \mathbf{Y}\mathbf{Y}^\top), \quad (4)$$

where $\mathbf{P} = (1-p)\mathbf{Y}\mathbf{Y}^\top$, $\mathbf{Q} = (1-p)^2\mathbf{Y}\mathbf{Y}^\top + p(1-p)\delta(\mathbf{Y}\mathbf{Y}^\top)$, and $\delta(\cdot)$ denotes the operation that sets all the entries except the diagonal of a matrix to zero. Then Eq. 4 can be considered as the regularization term for $\mathbf{B}$.

Finally, by integrating Eq. 2 and the regularization terms for $\mathbf{W}$, $\mathbf{U}$, and $\mathbf{B}$ discussed above, the proposed method can be formulated by minimizing:

$$\min_{\mathbf{W},\mathbf{U},\mathbf{B}} \frac{1}{n}\sum_{i=1}^{n}(\|\mathbf{B}\mathbf{y}_i - \mathbf{W}\mathbf{v}_i\|^2 + \|\mathbf{B}\mathbf{y}_i - \mathbf{U}\mathbf{t}_i\|^2) + \quad (5)$$
$$\lambda\|\mathbf{W}\|_2^2 + \mu\|\mathbf{U}\|_2^2 + \xi r(\mathbf{B}),$$

where λ, μ and ξ are the penalty coefficients of the regularization terms of each parameter.

2.2 Optimization

The minimization problem in Eq. 5 can be efficiently solved using block-coordinate descent algorithm by alternatively optimizing each variable by fixing the others. The detailed derivation for optimizing Eq. 5 is provided in supplementary. Here we depict the general optimization procedure of the parameters in each iteration as follows.

When fixing $\mathbf{U}$ and $\mathbf{B}$, optimizing $\mathbf{W}$ becomes a standard ridge regression problem that can be solved in closed-form:

$$\mathbf{W} = \mathbf{B}\mathbf{Y}\mathbf{V}^\top(\mathbf{V}\mathbf{V}^\top + n\lambda I)^{-1}. \quad (6)$$

Similarly, with fixed $\mathbf{W}$ and $\mathbf{B}$, $\mathbf{U}$ can also be optimized by solving the ridge regression problem, as:

$$\mathbf{U} = \mathbf{B}\mathbf{Y}\mathbf{T}^\top(\mathbf{T}\mathbf{T}^\top + n\mu I)^{-1}. \quad (7)$$

When fixing $\mathbf{W}$ and $\mathbf{U}$, $\mathbf{B}$ can be solved in closed-form as ordinary least squares:

$$\mathbf{B} = (\mathbf{W}\mathbf{V}\mathbf{Y}^\top + \mathbf{U}\mathbf{T}\mathbf{Y}^\top + \gamma\mathbf{P})[\mathbf{Y}\mathbf{Y}^\top + \gamma\mathbf{Q}]^{-1}, \quad (8)$$

where $\mathbf{P}$ and $\mathbf{Q}$ can be computed analytically with a predefined p. In practice, we can vary p in range $(0, 1)$ and choose appropriate value of p according to the resulting performance of label completion and cross-modal retrieval.

In summary, the objective function in Eq. 5 can be solved by updating variables $\mathbf{W}$, $\mathbf{U}$, and $\mathbf{B}$ iteratively until either convergence or a predefined number of iterations is reached. In each iteration, the time complexity is about $\mathcal{O}(kc^2n)$, where $k = \max\{m, d, c\}$ and usually we have $k \ll n$. Thus the training time complexity is linear to the size of training set, hence it is very efficient and scalable for large-scale training data.

2.3 Test Phase for Out-of-Instance

During the test phase, for a new test instance x' that contains features of one modality, i.e. $x' = (\mathbf{v}',)$ or $x' = (\mathbf{t}',)$, its subspace representation $\mathbf{z}'$ is computed as $\mathbf{z}' = \mathbf{W}\mathbf{v}'$ or $\mathbf{z}' = \mathbf{U}\mathbf{t}'$, respectively. For the the label completion task, we select the top-K labels that have top-ranked values in $\mathbf{z}'_i$ as the predicted label set of x', For the cross-modal retrieval task, we take $\mathbf{z}'$ as the query instance of one modality (e.g., image modality) to retrieve in the database of the other modality (text modality).

3. EXPERIMENTS

3.1 Experimental Setting

We apply the proposed method to both label completion and cross-modal retrieval tasks on two large-scale multimodal datasets: NUS-WIDE [3] and Flickr30K [32]. The NUS-WIDE originally containing 269,648 instances, with each being an image with its associated textual tags. Each instance is manually annotated with at least one of 81 labels. For each instance, its image is represented as a 500-D bag-of-words (BOW) SIFT feature vector and its text as a binary tagging vector w.r.t. the top 1,000 most frequent tags. In our experiment, we take 267,613 instances as database and the remaining 2,035 instances as query. The Flickr30K consists of 31,783 instances, with each being an image with associated five textual sentences. In [10], the most frequent 350 keywords from the sentences in Flickr30K are selected as the labels, hence each instance is naturally assigned at least one of 350 labels. For each instance, its image is represented as a 4,096-D convolutional neural networks (CNN) feature extracted by Decaf model [6], and its text as a tf-idf-weighted tagging vector w.r.t the top 3,000 most frequent words that are crawled in [10]. In our experiment, we take 28,917 instances as database and the remaining 2,866 instances as query. It is worth mention that, for each instance in the two datasets, the average number of labels are 2.4 and 3.1, respectively, which is indeed incomplete.

To reduce the computational cost during training, we randomly select 10,000 instances from the database of each dataset and then apply the learned model to the other instances in database. Finally, the label completion task is performed on the query instances, and the cross-modal retrieval task is conducted between the query and database instances. In all experiments, the parameters λ, μ and ξ of the proposed model are empirically set to 10^{-5}, 10^{-5} and 10^{-2}, respectively. The experiments are conducted on a desktop which has 4-core 3.3GHz CPUs with 16GB RAM. It is worth noting that training the proposed method is highly efficient on the two large-scale datasets. For example, the training time of the proposed method on Flickr30K is about 10 minutes when using the high dimensional features of image and text modalities.

3.2 Results of Label Completion Task

We first compare the proposed method with several related methods on the label completion task. These methods are divided into three groups: 1) traditional image annotation methods using unimodal image for label prediction: JEC [17] and TagProp [11]; 2) label completion methods using unimodal image: Fasttag [2] and LSR [15]; 3) the cross-modal retrieval methods LCFS [24] and ml-CCA [18], which are capable to use multimodal data of image and text for label prediction. We implement the fast version of ml-CCA according to the instruction its original paper. We use the standard measures that are used in both traditional image annotation and label completion tasks, including average precision per label (P), average recall per label (R) and the number of labels that are recalled (N_+). For all the metrics, larger numerical value indicates better performance.

Table 1: Comparison of the proposed method with its counterparts on label completion task.

Dataset	NUS-WIDE			Flickr30K		
Method	P	R	N+	P	R	N+
JEC	0.031	0.102	54	0.263	0.312	249
TagProp	0.053	0.127	47	0.312	0.368	267
Fasttag	0.068	0.134	45	0.321	0.387	299
LSR	0.083	0.152	63	0.377	0.392	315
LCFS	0.398	0.478	69	0.417	0.491	319
ml-CCA	0.464	0.476	72	0.413	0.511	322
Proposed	**0.491**	**0.512**	**73**	**0.430**	**0.533**	**336**

Table 1 shows the overall results of different approaches on the two datasets. From Table 1, we have the following observations. Firstly, the multi-modal approaches (e.g., LCFS, ml-CCA) generally outperform the unimodal ones (e.g., JEC, TagProp, Fasttag, and LSR), since the additional text modality is helpful for label prediction. In addition, the visual feature of the NUS-WIDE dataset is less effective for the approaches that using unimodal images for label prediction. Secondly, these two datasets really suffer from missing labels, and the label completion approaches Fasttag and LSR improve the label prediction results of traditional image annotation methods JEC and TagProp. Moreover, the proposed method also outperforms the multi-modal approaches LCFS and ml-CCA that neglect missing labels. Thirdly, the proposed method achieves the best performance for label prediction on both datasets, showing the superiority of the proposed cross-modal co-regularized learning framework on capturing the hidden label correlation from multi-modal features of image and text.

3.3 Results of Cross-modal Retrieval Task

In this subsection, we evaluate the proposed method on cross-modal retrieval task. Specifically, we consider two standard scenarios (sub-tasks): Img2Txt and Txt2Img. We compare the proposed method with several subspace learning methods, including unsupervised ones: CCA [19], 3-view CCA [9] and supervised ones that can handle multi-label assumption: LCFS [24], CDL [28] ml-CCA [18]. We adopt the same evaluation metric of mean average precision (MAP) that is widely used in these methods for the two sub-tasks.

Table 2: Comparison of the proposed method with its counterparts on cross-modal retrieval task.

Dataset	NUS-WIDE		Flickr30K	
Method	Img2Txt	Txt2Img	Img2Txt	Img2Txt
CCA	0.265	0.278	0.228	0.245
3-view CCA	0.294	0.312	0.262	0.288
LCFS	0.349	0.363	0.271	0.296
CDL	0.358	0.366	0.293	0.301
ml-CCA	0.362	0.383	0.307	0.311
Proposed	**0.374**	**0.399**	**0.321**	**0.328**

Table 2 shows the MAP scores of all the methods on the two sub-tasks. We can observe that the results of unsupervised methods CCA and 3-view CCA are worse than the proposed method and other supervised ones LCFS, CDL, and ml-CCA. The main reason is that the unsupervised methods only use pairwise information of image and text modalities, ignoring the valuable label information. Furthermore, the proposed method outperforms the other supervised methods LCFS, CDL, and ml-CCA, indicating that it benefits from the label completion to learn more discriminant subspace for cross-modal matching.

4. CONCLUSION

In this paper, we studied the problem of cross-modal retrieval on multi-modal data with incomplete labels. We proposed a novel method that can simultaneously perform label completion and cross-modal retrieval. To achieve this, we introduced a common subspace as the ideally complete labels, which was jointly learned from two aspects: 1) linear projection from the modality-specific features; and 2) enriching mapping from the incomplete labels. This scheme ensures that the learned subspace was discriminant to account for the complete label information. We developed an efficient iterative algorithm for solving the optimization problem. Experiments on two large-scale datasets validated the advantages of the proposed method on both label completion and cross-modal retrieval tasks.

5. ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Project 61572108 and Project 61502081, the National Thousand-Young-Talents Program of China, and the Fundamental Research Funds for the Central Universities under Project ZYGX2014Z007, Project ZYGX2015J055 and Project ZYGX2016KYQD114.

6. REFERENCES

[1] D. M. Blei and M. I. Jordan. Modeling annotated data. In *ACM SIGIR*, pages 127–134, 2003.

[2] M. Chen, A. Zheng, and K. Weinberger. Fast image tagging. In *ICML*, pages 1274–1282, 2013.

[3] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *CIVR*, 2009.

[4] C. Deng, X. Tang, J. Yan, W. Liu, and X. Gao. Discriminative dictionary learning with common label alignment for cross-modal retrieval. *TMM*, 18(2):208–218, 2016.

[5] D. S. Doermann. The indexing and retrieval of document images: A survey. *Computer Vision and Image Understanding*, 70(3):287–298, 1998.

[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, volume 32, pages 647–655, 2014.

[7] F. Feng, X. Wang, and R. Li. Cross-modal retrieval with correspondence autoencoder. In *ACM MM*, pages 7–16, 2014.

[8] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *ICCV*, volume 2, pages 1816–1823, 2005.

[9] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106:210–233, 2014.

[10] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*, volume 8692, pages 529–545, 2014.

[11] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *CVPR*, pages 309 –316, 2009.

[12] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664, 2004.

[13] S. Hwang and K. Grauman. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *IJCV*, 100:134–153, 2012.

[14] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan. Learning consistent feature representation for cross-modal multimedia retrieval. *TMM*, 17(3):370–381, 2015.

[15] Z. Lin, G. Ding, M. Hu, J. Wang, and X. Ye. Image tag completion via image-specific and tag-specific linear sparse reconstructions. In *CVPR*, pages 1618–1625, 2013.

[16] Y. Liu, D. Zhang, G. Lu, and W. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.

[17] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV*, 2008.

[18] V. Ranjan, N. Rasiwasia, and C. V. Jawahar. Multi-label cross-modal retrieval. In *ICCV*, pages 4094–4102, 2015.

[19] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, 2010.

[20] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *Proc. SLSFS*, pages 34–51, 2006.

[21] F. Shen, W. Liu, S. Zhang, Y. Yang, and H. Tao Shen. Learning binary codes for maximum inner product search. In *ICCV*, pages 4148–4156, 2015.

[22] F. Shen, C. Shen, W. Liu, and H. T. Shen. Supervised discrete hashing. In *CVPR*, pages 37–45, 2015.

[23] H. Wang, S. Liu, and L.-T. Chia. Does ontology help in image retrieval?: a comparison between keyword, text ontology and multi-modality ontology approaches. In *ACM MM*, pages 109–112, 2006.

[24] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. Learning coupled feature spaces for cross-modal matching. In *ICCV*, pages 2088–2095, 2013.

[25] W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. In *ICML*, 2015.

[26] B. Wu, S. Lyu, and B. Ghanem. ML-MG: multi-label learning with missing labels using a mixed graph. In *ICCV*, pages 4157–4165, 2015.

[27] X. Xu, A. Shimada, H. Nagahara, and R. Taniguchi. Learning multi-task local metrics for image annotation. *Multimedia Tools Appl.*, 75(4):2203–2231, 2016.

[28] 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 MM*, pages 847–850, 2015.

[29] Y. Yang, Y. Yang, and H. T. Shen. Effective transfer tagging from image to video. *ACM Trans. Multimedia Comput. Commun. Appl.*, 9(2):1–20, 2013.

[30] Y. Yang, Z. Zha, Y. Gao, X. Zhu, and T. Chua. Exploiting web images for semantic video indexing via robust sample-specific loss. *TMM*, 16(6):1677–1689, 2014.

[31] Y. Yang, H. Zhang, M. Zhang, F. Shen, and X. Li. Visual coding in a semantic hierarchy. In *ACM MM*, pages 59–68, 2015.

[32] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.