ENHANCED ISOMORPHIC SEMANTIC REPRESENTATION FOR CROSS-MEDIA RETRIEVAL

Ting liu, Yao Zhao, Shikui Wei, Yunchao Wei and Lixin Liao*

Institute of Information Science, Beijing Jiaotong University, Beijing, 100044, China
Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, 100044, China
16112055,yzhao,shkwei@bjtu.edu.cn, wychao1987@gmail.com, 16112056@bjtu.edu.cn

ABSTRACT

Nowadays cross-media retrieval is an useful technology that helps people find expected information from the huge amount of multimodal data more efficiently. A common cross-media retrieval framework is first to map features of different modalities into an isomorphic semantic space so that the similarity between heterogeneous data can be measured. For most of semantic space based methods, the mapping mechanism from original to semantic space of each modality is optimized independently, yet the more discriminative characteristic of a certain modality is not taken into account. In this paper, we propose a deep framework which introduces a latent embedding layer to learn joint parameters to obtain semantically meaningful representations of images and texts. Specifically, the discriminative characteristic embedded in the textual modality can be transferred to images through the latent embedding layer and joint parameters to enhance the consistency between semantic representations. Extensive experiments on the three popular publicly available datasets well demonstrate the superiority of the proposed method, which achieves the new state-of-the-arts.

Index Terms— cross-media retrieval, deep learning, subspace learning

1. INTRODUCTION

With the development of information technology, a large number of data (*e.g.* image, text and video) is generating on the Internet. These data of different modalities often co-occur to describe the similar content. For instance, an image surrounding with texts on a web page usually describes the same object of event. In recent years, cross-media retrieval, *e.g.* using image to retrieve text or using text to retrieve image, has attracted much attention. However, the problem of cross-media retrieval is very challenging and keeps still open.

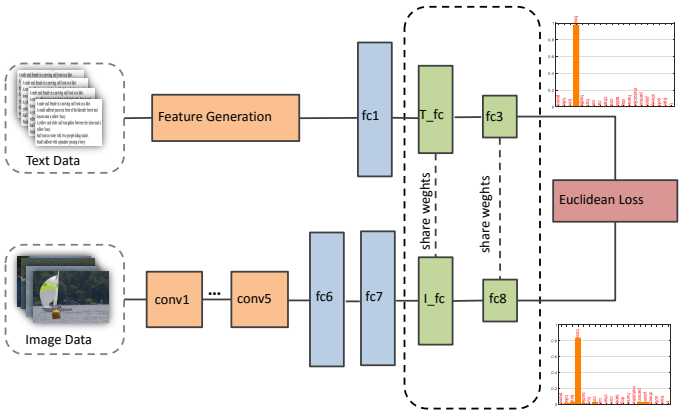

Fig. 1. Illustration of the proposed framework. A novel latent embedding layer and shared weights scheme are introduced between two sub-networks so that the discriminative semantic characteristic embedded in the textual modality can be leveraged for optimizing the image subnetwork.

One of great challenges is so-called heterogeneous gap among multi-modal data. That is, the instances of multi-modal data are represented by totally different feature spaces. For example, feature representations of image and text are always with different dimensions, thus content similarity between image and text cannot be directly estimated. To address this problem, one of popular solutions is to map feature representations of different modalities into an isomorphic space, in which the similarity between different modalities can be directly measured.

During the past few years, numerous methods [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12] have been proposed to address the cross-media retrieval task by learning an optimal isomorphic representation for multiple modalities. As a popular common space learning method, Canonical Correlation Analysis (CCA) [1] is usually employed to find a pair of mapping matrices to maximize the correlations between two kind of feature representations. Based on CCA, many extensions [4, 5, 6, 7] have been developed for cross-media

This work was supported in part by National Natural Science Foundation of China (No.61572065, No.61532005), Joint Fund of Ministry of Education of China and China Mobile (No.MCM20160102), and Fundamental Research Funds for the Central Universities (No.2015JBZ002).

*S. Wei is corresponding author.

978-1-5090-6067-2/17/$31.00

retrieval. Sharma *et al.* [4] proposed a generic framework, called Generalized Multiview Analysis, to map feature representations in different modality spaces into an isomorphic (non)linear space. Gong *et al.* [5] proposed a three-view CCA method by introducing a semantic view to produce a better separation for multi-modal data belonging to different categories in the learned common space. Rasiwasia *et al.*l [6] proposed a cluster CCA method to learn discriminant isomorphic representations that maximize the correlation between two modalities while distinguishing the different categories. Wei *et al.* [7] proposed a modality-independent cross-media retrieval method, which learns two couples of projections for different retrieval tasks. For each couple of projections, the correlation between two modalities and the linear regression from one modality space to the common space are jointly optimized to learn the isomorphic representation.

Beyond the CCA-based methods, Rasiwasia *et al.* [3] proposed a Semantic Matching (SM) method to address the cross-media retrieval problem. In particular, SM is to generate probability distribution over classes as the semantic features. In this way, the correspondences between the data of different modalities can be built naturally. Following the same motivation, a deep version of SM is proposed by Wei *et al.*. [8] for learning better common semantic representations. Peng *et al.* [13] proposed a cross-media multiple deep network to exploit the complex cross-media correlation by hierarchical learning. Recently, Liu *et al.* [14] proposes a new evaluation protocol for cross-media retrieval which better fits the real-word applications. However, for each modality, the semantic representation is optimized independently and the correlations between two modalities are not taken into account. It should be noted that the feature representation of text is usually with more discriminative characteristic compared with that of image. Thus, if this characteristic can be balanced with images during the optimization process, the quality of semantic representations for images may be improved, which will benefit the cross-media retrieval.

Following this observation, in this paper, we propose a deep framework to learn the enhanced isomorphic semantic representations of images and texts for cross-media retrieval as shown in Fig. 1. Within the deep framework, two subnetworks are embedded to map images and texts from their original feature spaces into an isomorphic semantic space. To enhance the semantic representations of images, a novel latent embedding layer is introduced between two sub-networks so that the discriminative semantic characteristic embedded in the textual modality can be leveraged for optimizing the image subnetwork. The main contributions of this paper can be summarized as follows:

- An novel latent embedding layer is embedded in the proposed framework, so that the discriminative characteristic of texts can be utilized to enhance the semantic representations of images with the joint parameters.
- Extensive experiments well demonstrate the superiority of the proposed framework. Specifically, the mAP scores on Wikipedia and Pascal Sentence can reach 44.1% and 49.5%, which are the new state-of-the-arts.

2. THE PROPOSED METHOD

The whole deep framework includes two sub-networks, which are utilized to extract isomorphic semantic representations for images and texts. And a novel latent embedding layer is introduced between two sub-networks, in which its weights are shared by two sub-networks. For the following descriptions, we denote the sub-networks for image and text as I_Net and T_Net, respectively.

2.1. Image Net

Training a CNN model with high performance needs a large amount of label data, while it is very difficult for many recognition tasks to get these labelled data. During the past years, many works have proved that transferring the CNN models pre-trained on large datasets to deal with some specific tasks is a good choice. Therefore, for the I_Net, we employ AlexNet [15] as the basic architecture, whose parameters are pre-trained on ImageNet [16]. Since directly applying the pre-trained CNN model to extract visual features is not the best strategy, we employ the images from target dataset to fine-tune the pre-trained parameters. As shown in Fig. 1, the basic AlexNet contains five convolution layers (short as conv) and three fully-connected layers (short as fc). However, it is not easy to transfer the discriminative characteristic of texts into the original network model. To facilitate the process, we insert a fully connected layer I_fc into the middle between fc7 layer and fc8 layer.

2.2. Text Net

Since the text feature is usually more discriminative than image, it is instinct that the relationship between text features and their semantics can be more easily built. Hence, for the T_Net, we adopt a neural network with three fully-connected layers to map text features from original feature space into semantic space. Similar to the fully-connected layers in CNN, ReLU is utilized as the nonlinear activation function for each fully-connected layer.

2.3. Enhanced Isomorphic Semantic Representation

Suppose there are n pairs of image and text for training, *i.e.*, $\mathcal{G} = \{(\boldsymbol{x}_i, \boldsymbol{t}_i)\}_{i=1}^{n}$, where $\boldsymbol{x}_i$ and $\boldsymbol{t}_i$ denote image and text, respectively. Each pair is associated with a label vector $\boldsymbol{y}_i$ of c components, where c is the number of classes. We set the $j^{th}(j = 1, \cdots, c)$ component of $\boldsymbol{y}_i$ as 1 if $\boldsymbol{x}_i$ and $\boldsymbol{t}_i$ are labeled with the j^{th} class and 0 otherwise.

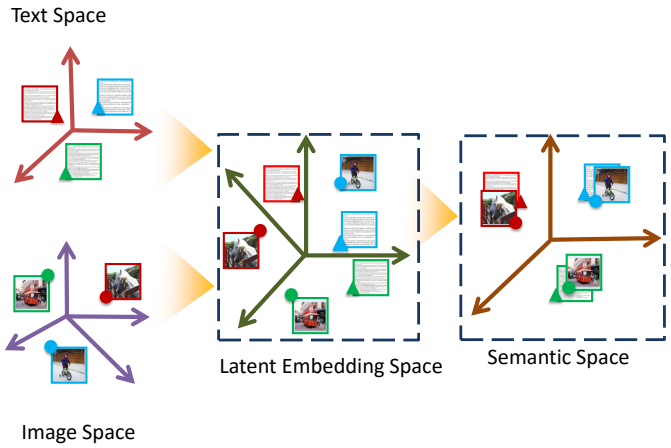

Fig. 2. Illustration of joint optimization scheme. For each modality in the heterogeneous space, we map the feature representations of images and texts from the original space into latent embedding space, in which the shared weights of two subnetworks can be jointly learned.

In order to obtain isomorphic semantic representations for images and texts, we consider a metric to minimize the distance between semantic representation of image (or text) and it corresponding label vector. Formally, we denote the parameters of two sub-networks as W_I and W_T. Then the objective function to be minimized is defined as follows:

$$\mathcal{L}_S = \sum_{i=1}^{n}(\left\| f^{(I)}_{W_I}(\boldsymbol{x}_i) - \boldsymbol{y}_i \right\|^2 + \left\| f^{(T)}_{W_T}(\boldsymbol{t}_i) - \boldsymbol{y}_i \right\|^2) \quad (1)$$

where $f^{(I)}_{W_I}$ and $f^{(T)}_{W_T}$ are the outputs with softmax operation from the last fully connected layers of I_net and T_net respectively, which are the semantic representations of $\boldsymbol{x}_i$ and $\boldsymbol{t}_i$. In essence, however, the parameters of T_net and I_Net are learned independently, while they are combined into one objective function Eqn.(1). In this way, the semantic representations for pairs of image and text cannot be well consistent. As a result, the semantic relationship between image and text cannot be aligned well, leading to bad cross-media retrieval accuracy. To address this issue, we attempt to bridge the gap between two subnetworks and make the discriminative capability of both modalities more powerful. According to our observation, the discriminative capability of textual modality is more powerful than the image modality. Therefore, an intuitive idea is to learn joint parameters for both modalities when feeding into the last layer, as shown in Fig. 2. Towards this end, we propose a novel latent embedding layer to enhance the consistency of semantic representations for pairs of image and text. As shown in Fig. 1, the latent embedding layer relays on two feature representations as inputs, *i.e.*, I-fc and T-fc. The output representations for I-fc and T-fc are with the same dimension in the latent space. We denote the learned representations of I-fc and T-fc as $\phi(\boldsymbol{x}_i)$ and $\phi(\boldsymbol{t}_i)$, respectively. To jointly learn the parameters for image and text, we share weights in the last fully-connected layers of two subnetworks, which maps the features from latent embedding space into semantic space, as shown in Fig. 2. Formally, the proposed method can be formulated into the following optimization function:

$$\min_{W} \mathcal{L} = \sum_{i=1}^{n}(\|W\phi(\boldsymbol{x}_i) - \boldsymbol{y}_i\|^2 + \|W\phi(\boldsymbol{t}_i) - \boldsymbol{y}_i\|^2) \quad (2)$$

It is worth noticing that the parameters of W are embedded in different layers of the deep framework and can be optimized with Stochastic Gradient Descent (SGD).

3. DATASETS AND IMPLEMENTATION DETAILS

In order to evaluate the effectiveness of proposed method, we conduct a series of experiments on three popular publicly available datasets, *i.e.*, Wikipedia [3], Pascal Sentence [17] and Pascal VOC 2007 [18].

3.1. Datasets

Wikipedia: This dataset contains 2,866 image-text pairs from ten classes in total, which is randomly split into a training set and a test set with 2,173 and 693 pairs, respectively. For text representation, we first employ some natural language processing techniques (including stop word removing, stemming and lemmatization) to preprocess the text documents. Then, Latent Dirichlet allocation (LDA) [19] is utilized to extract text features under 100 topics.

Pascal Sentence: This dataset contains 1,000 image-text pairs from 20 classes (50 for each class). We adopt the same splits as reported in [7], *i.e.*, 30 pairs from each class for training and the rest for testing. For text representation, some preprocessing techniques are first conducted and LDA is utilized to extract text features under 100 topics.

Pascal VOC 2007: This dataset contains 9963 image-text pairs from 20 classes. The data has been split into 5011 pairs for training/validation and 4952 pairs for testing. The 798 dimensional tag ranking feature provided by [20] is utilized as the text feature.

In our experiments, Euclidean distance is used to estimate the similarity between text features and image features. Retrieval performance is evaluated by mean Average Precision (mAP), which is one of the standard information retrieval metrics. Average precision is used to evaluate the accuracy of a single query, which is the average of precisions computed at the point of each of the test samples in the ranked sequence. The average precisions is defined as:

$$AP = \frac{\sum_{r=1}^{N} P(r)rel(r)}{\sum_{r=1}^{N} rel(r)} \quad (3)$$

where N is the size of test dataset, $rel(r) = 1$ if the result at rank r is relevant, otherwise $rel(r) = 0$. $P(r)$ is the precision of the result at rank r. The average of all queries is mAP. Note that Pascal VOC 2007 is multi-label datasets. Following [21], we treat one result as relevant if it share at least one class label with the query.

3.2. Implementation Details

The proposed framework employs AlexNet [15] as the basic architecture of the I_Net, which is pre-trained on ILSVRC 2012 [16] classification dataset. We make a change on the AlexNet, *i.e.*, adding a new latent embedding fully-connected layer into the middle of fc7 layer and fc8 layer. The architecture of T_Net is constructed with three fully-connected layers (4096-l-c, where l is the dimension of latent embedding layer, and c is the number of classes for a specific dataset). Drop out and ReLU are added likewise.

All the newly added layers in both I_Net and T_Net are randomly initialized with zero-mean Gaussian distributions with standard deviations of 0.01. All the training and testing images are normalized to 256 $\times$ 256. During the training stage, we use 256 image-text pairs per mini-batch. In order to expand the training dataset, training images are horizontally flipped with a probability of 0.5 and randomly cropped at 227 by 227. The initial learning rates of newly added layers and pre-trained layers are set as 0.01 and 0.001, respectively. The learning rates of all layers decrease to one tenth of the current rates at the half of the total iterations. We run SGD for 300 iterations in total to train the network parameters for Pascal Sentence dataset, 400 iterations for Wikipedia dataset, 2000 iterations for Pascal VOC 2007 dataset. The deep framework is trained based on a NVIDIA Tesla K40.

4. EXPERIMENTS

4.1. Ablation Studies

To validate the effectiveness of the experimental settings, we conduct ablation studies on Pascal Sentence, Wikipedia and Pascal VOC.

4.1.1. The Dimension of Latent Embedding Layer

Different size of the fully-connected layer will directly effect the number of learned parameters and the effectiveness of the network. Therefore, we should determine the size of the latent embedding layer, before we add a new latent embedding layer.

To evaluate the effect of the dimension on the latent embedding layer(I-fc or T-fc), we conduct a series of cross-media retrieval experiments by varying the dimension of latent embedding layer on three datasets. In our experiments, we vary the dimension from 64 to 1024, stepping by a power of two each time. From the experimental result in Table 1, the optimal dimensions are different for different datasets. The main reason lies in that the size and class number of different datasets are various. In the following experiments, we choose 128 as the dimension of the latent embedding layer for Pascal Sentence dataset, and choose 512 as the dimension for Wikipedia dataset and Pascal VOC 2007 dataset.

Table 1. Comparisons of mAP scores (%) by varying the dimension of latent embedding layer.

Dataset	Dim	Image Query	Text Query	Average
Pascal Sentence	64	40.9	50.4	45.7
	128	46.5	**52.4**	**49.5**
	256	47.0	51.5	49.3
	512	**47.9**	50.1	49.0
	1024	46.9	47.9	47.4
Wikipedia	64	40.3	44.7	42.5
	128	43.0	**45.1**	44.1
	256	43.5	44.9	44.2
	512	**44.3**	45.0	**44.6**
	1024	44.0	44.4	44.2
Pascal VOC 2007	64	77.9	75.4	76.6
	128	83.9	78.5	81.2
	256	84.9	79.2	82.0
	512	85.2	**79.3**	**82.3**
	1024	**85.5**	79.1	**82.3**

4.1.2. Effectiveness of the latent embedding space and joint parameters learning

Comparing with the method proposed in [7], we change the dimension of each layer in T_Net and add a new latent embedding layer. Therefore, the proposed method can jointly learn the parameters at the stage of mapping the representations in latent space into semantic space, which results in a more discriminative feature representation. In this section, we conduct several experiments to validate the effectiveness of the proposed new network.

Firstly, we validate the effectiveness of the latent embedding layer (O/LE) , compared with the original network (O). Secondly, we further validate the effectiveness of joint parameters learning based on the latent embedding layer (O/LE/J). As shown in Table 2, the proposed framework obtains consistent improvements compared with the original network and the original network with only the latent embedding layer. These results indicate that the joint parameter learning is vital for the new network architecture. In fact, it is reasonable because the joint parameter learning ensures that the image data and the text data have more similar distribution in the semantic space. In brief, the proposed approach can obtain a better isomorphic semantic representation for the image data and the text data, which will benefit the cross-media retrieval

Table 3. Comparisons of Cross-media Retrieval Performance

Dataset	Query	CCA [2]	SM [3]	T-V CCA [5]	GMLDA [4]	GMMFA [4]	DEEP-SM [8]	MDCR [7]	OURS
Wikipedia	Image	22.6	40.3	31.0	37.2	37.1	39.8	43.5	**44.3**
	Text	24.6	35.7	31.6	34.7	34.6	35.4	39.4	**45.0**
	Average	23.6	38.0	31.3	34.7	34.6	37.6	41.5	**44.6**
Pascal Sentence	Image	26.1	42.6	33.7	45.6	45.5	44.6	45.5	**46.5**
	Text	35.6	46.7	43.9	44.8	44.7	47.8	47.1	**52.4**
	Average	30.9	44.6	38.8	45.2	45.1	46.2	46.3	**49.5**
Pascal VOC 2007	Image	66.1	-	71.5	-	-	82.3	-	**85.2**
	Text	66.8	-	73.4	-	-	77.6	-	**79.3**
	Average	66.5	-	72.5	-	-	80.0	-	**82.3**

Table 2. Comparisons of mAP scores (%) to validate the effectiveness of latent embedding layer and joint parameters learning.

Dataset	Query	O	O/LE	O/LE/J
Pascal Sentence	Image	47.6	45.2	**46.5**
	Text	50.0	51.7	**52.4**
	Average	48.8	48.9	**49.5**
Wikipedia	Image	42.0	41.1	**43.0**
	Text	44.1	45.0	**45.1**
	Average	43.1	43.1	**44.1**
Pascal VOC 2007	Image	84.2	84.4	**85.2**
	Text	78.9	78.7	**79.3**
	Average	81.6	81.5	**82.3**

performance.

4.2. Comparison with State-of-the-art Methods

In order to further prove the effectiveness of our method, we conduct comparative experiments between the proposed method and other seven methods, *i.e.*, CCA, Semantic Matching (SM), Three-View CCA (T-V CCA), Generalized Multi-view Marginal Fisher Analysis (GMMFA), Generalized Multi-view Linear Discriminant Analysis (GMLDA), deep semantic match (DEEP-SM) and Modality-Dependent Cross-media Retrieval method (MDCR). In these experiments, the visual feature of one image is 4096-dimensional vector extracted by the pre-trained AlexNet, and the textual feature of one text is 100-dimensional LDA for both Wikipedia and Pascal Sentence.

The mAP scores of these methods are directly token from [7] and [8]. As shown in Table 3, the proposed method obtains significant and consistent improvements on all three datasets compared with other seven methods. These results further prove that the proposed method achieves the state-of-the-art level. For example, the mAP of the proposed method on Pascal Sentence improves 3.2% (46.3% *vs.* 49.5%), compared the best result in other seven methods. Note that SM, GMLDA and GMMFA schemes cannot fit the multi-label dataset, *i.e.*, Pascal VOC 2007. Therefore, their experimental results are not available. As expected, the proposed method outperforms all other methods on the dataset. For example, the proposed method improves 2.3% (80.0% *vs.* 82.3 %), on Pascal VOC 2007 dataset, compared to the best result in other three methods.

To illustrate the effectiveness in a more intuitive way, we demonstrate some examples of computed semantic representations for pairs of image and text on Pascal Sentence. As shown in Fig. 3, we can evidently mention that the bin on their own semantic class is dominantly higher than others. That is, the proposed method indeed captures the semantic information in both modalities.

5. CONCLUSION

This paper proposes a deep framework which introduces a latent embedding layer to learn joint parameters. By mapping the image and text representations from the latent space into semantic space with joint parameters, we can obtain more semantically meaningful representations for images and texts. Extensive experiments on three popular publicly available datasets demonstrate that the proposed method indeed enhances the isomorphic semantic representation and shows the superiority compared with state-of-the-art methods.

6. REFERENCES

[1] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[2] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis.," in *ICML*, 2013, pp. 1247–1255.

[3] N. Rasiwasia et al., "A new approach to cross-modal multimedia retrieval," in *ACM MM*, 2010, pp. 251–260.

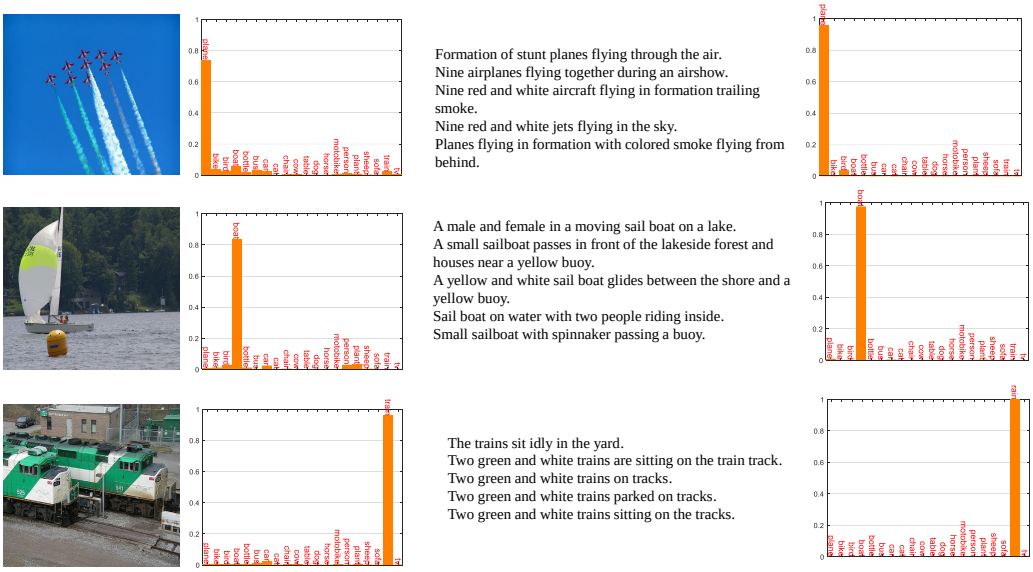

Fig. 3. Examples of computed semantic representations for pairs of image and text on Pascal Sentence.

[4] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *CVPR*, 2012, pp. 2160–2167.

[5] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *IJCV*, vol. 106, no. 2, pp. 210–233, 2014.

[6] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," in *AISTATS*, 2014, pp. 823–831.

[7] Y. Wei et al., "Modality-dependent cross-media retrieval," *TIST*, vol. 7, no. 4, pp. 57, 2016.

[8] Y. Wei et al., "Cross-modal retrieval with cnn visual features: A new baseline," *TCYB*, 2016.

[9] R. Liu, Y. Zhao, S. Wei, and Z. Zhu, "Cross-media hashing with centroid approaching," in *ICME*, 2015, pp. 1–6.

[10] S. Wei, Y. Wei, L. Zhang, Z. Zhu, and Y. Zhao, "Heterogeneous data alignment for cross-media computing," in *ICIMCS*, 2015, p. 84.

[11] Y. Wei, Y. Zhao, Z. Zhu, Y. Xiao, and S. Wei, "Learning a mid-level feature space for cross-media regularization," in *ICME*, 2014, pp. 1–6.

[12] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *TMM*, vol. 17, no. 3, pp. 370–381, 2015.

[13] Y. Peng, X. Huang, and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks," .

[14] R. Liu, Y. Zhao, L. Zheng, S. Wei, and Y. Yang, "A new evaluation protocol and benchmarking results for extendable cross-media retrieval," *arXiv preprint arXiv:1703.03567*, 2017.

[15] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[16] J. Deng et al., "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[17] C. Rashtchian, P. Young, Peter, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *ACL*, 2010, pp. 139–147.

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 1, pp. 993–1022, 2003.

[20] S. J. Hwang and K. Grauman, "Accounting for the relative importance of objects in image retrieval," in *BMVC*, 2010, vol. 1, p. 5.

[21] Y. Wei et al., "Hcp: A flexible cnn framework for multi-label image classification," *TPAMI*, vol. 38, no. 9, pp. 1901–1907, 2016.