A Novel Cross Modal Hashing Algorithm Based on Multi-modal Deep Learning

Wen Qu[1(✉)], Daling Wang[1,2], Shi Feng[1,2], Yifei Zhang[1,2], and Ge Yu[1,2]

[1] School of Information Science and Engineering,
Northeastern University, Shenyang, China
quwen@research.neu.edu.cn

[2] Key Laboratory of Medical Image Computing,
Northeastern University, Ministry of Education, Shenyang, China
{wangdaling,fengshi,zhangyifei,yuge}@ise.neu.edu.cn

Abstract. With the popularity of multi-modal data on Web, cross media retrieval has become a hot research topic. Existing cross modal hash methods assume that there is a latent space shared by multi-modal features, and embed the heterogeneous data into a joint abstraction space by linear projections. However, these approaches are sensitive to the noise of data, and unable to make use of unlabelled data and multi-modal data with missing values in the real-world applications. To address these challenges, in this paper, we propose a novel Multi-modal Deep Learning based Hashing (MDLH) algorithm. In particular, MDLH adopts deep neural network to encode heterogeneous features into a compact common representation and learn the hash functions based on the common representation. The parameters of the whole model are fine-tuned in supervised training stage. Experiments on two standard datasets show that our method achieves more effective results than other methods in cross modal retrieval.

1 Introduction

As the popularity of social media in the Web 2.0, the amount of multi-modal data increases dramatically in recent years. For example, photos are usually associated with captions and tags, videos contain visual and audio signals, and tweets often consist of text, images and videos. At the same time, when users acquire and search through the Internet, they also want to get a comprehensive result consisting of multiple media types. The traditional information retrieval system only uses text as query input, so most information systems provide the image and video retrieval based on text queries. With the rapid development of the mobile equipment such as telephone and flat computer, users may perform queries using image, audio and videos other than text. There is an emerging need to retrieve and search similar or relevant data entities from multiple modals. To make the

W. Qu—This work is supported by the National Natural Science Foundation of China under Grant No. 61402091, 61370074 and the Fundamental Research Funds for the Central Universities of China under Grant No. N140404012.

X. Zhang et al. (Eds.): SMP 2015, CCIS 568, pp. 156–167, 2015.
DOI: 10.1007/978-981-10-0080-5_14

system possible for handling large amount of multimedia data, hashing based methods have attracted increasing attentions due to the advantages in reducing both the computational cost and storage. A lot of work extended uni-modal hashing into multi-modal setting [23]. Cross modal hashing maps data of different modalities into the hamming space, in which the distance of similar objects to be small. In the hamming space, all data are represented as hash codes and can be searched quickly even for the databases with millions of data. Most previous cross modal hashing methods follow the assumption that multi-modal data used for training are available in all the multiple modals and contain the same 'semantic object'. So these works can not make use of unlabeled data or multi-modal data with missing values. In realistic applications, the data in the Internet is typically very noisy and may have missing modals. For example, the image and text of a tweet may contain different semantics at all. Furthermore, given a system supporting cross modal retrieval including text, image and audio, if the data generated by users only contain text and image, they cannot be used for modeling relationship among the three modals. Most previous works represent multi-modal data through clustering [25], dictionary learning [22], which build the corresponding maps pair-wise. When a new modal is added to the system, the relationship of the new modal with each existing modal has to be learned again. To address these problems, in this paper we propose a Multi-modal Deep Learning based Hashing (MDLH) algorithm, which learn the common feature space of different modalities using deep neural network. The multi-modal deep learning can learn compact and robust 'semantic' representation of multi-modal data, which is able to handle the noise and the missing modals of the data. The experiments on two realistic datasets show that the proposed method can realize cross modal hashing effectively. The rest of the paper is organized as follows: In Sect. 2, we review the related work. Section 3 elaborates the method proposed in this paper. In Sect. 4, we demonstrate the use of our approach for cross modal retrieval and the experimental results. Finally, we conclude the work in Sect. 5.

2 Related Work

The work involves with cross modal hashing and multi-modal deep learning, which will be reviewed in the following subsections.

2.1 Cross Modal Hashing

Hashing index can be categories into uni-modal hashing, multi-modal hashing, and cross modal hashing. In the work about uni-modal hashing, the most well known methods are local sensitive hashing [5] and spectral hashing [20]. Multi-modal hashing compares the multi-modal features of data, and returns the search results of each modal. For example, when retrieving an image according to multi-modal (color, SIFT, BOW) descriptors, the multi-modal hashing projects each feature into the hamming space and combines the multiple results together. Cross modal hashing focuses on analyzing the relationship between modalities and provides cross modal query. For example, given the color feature of an image as the input, the system returns the results according to SIFT descriptor. Here the modal means feature or media type. So cross modal means cross feature or cross media. The existing uni-modal data hashing includes two steps: First, project the original data into low-dimensional space. Then, quantize the new representation into hash codes. Under the unsupervised situation, many embedding methods have been proposed, such as random projection [5], spectral decomposition [20]. Similarly, multi-modal data hashing includes two steps with more restrictions. Bronstein [3] proposed the first cross modal hashing model CMSSH. Given two modals, CMSSH learned two groups of hash functions that made the similar data (in different modals) have smaller distance in the hamming space while dissimilar data (in different modals) have larger distance in the hamming space. CMSSH kept the relationship between data in different modals but ignored the similarity in same modal. Kumar [10] extended the spectral hashing into multimodal setting and proposed CVH, which minimized the distance of similar data both in the same modal and the different modals. MLBE [23] used probability generative model to represent the data, and the latent factors learned were used as the hash codes. There is no independent restrict of hash codes so the hash codes may have high redundancy. Yu [22] adopted dictionary learning to represent data in different modals, and learned the hash function based on sparse codes. The dictionaries of different modals were connected through the coupled dictionary space. IMVH [8] kept both intra similarity and inter similarity of the data. Song et al. [17] proposed Inter-Media Hashing which used a set of corresponding image and text as the inter media to learn the relations of multiple modals.

2.2 Multi-modal Deep Learning

Deep learning builds a layer network structure to simulate the human brain, and learn representations for data from bottom to up. Each layer of the network corresponds to a representation. Recently, deep learning is widely used in many applications and achieves impressive results, including speech recognition [6], face recognition, image classification [9] and object recognition etc. The representative deep learning including Deep Belief Networks [1], Auto Encoder, Stacked Denoising Autoencoder, Deep Boltzmann Machine and Deep Energy Model. Ngiam [12] used DBM to learn the cross modal representation of video and audio data, and rebuilt the data of missed modality. Srivastava [18] proposed a Deep Belief Network to learn representation of the multi-modal data. Sohn [16] proposed an improved multi-modal deep learning model. These works focus on solving the data rebuilding problem when part of the modality is missing. Our work focuses on learning the relations between different modals and proposing a semantic and common representation of multi-modal data. The work most similar to ours is Wu [21], in which the deep learning is used to learn the optimal combination of different modalities. Different from their work, we focus on learning a common representation of multi-modal data using deep neural network.

3 The Multi-modal Deep Learning Based Hashing Algorithm Methodology

In this section, we present the MDLH algorithm in detail. Figure 1 is the framework of our method. First, the multi-modal features of multi-modal data are extracted as inputs. Then, we use multi-modal deep learning method to learn the common representation for them. Finally, the hash function of each modality is used to map the data into the hamming space. In the following, the notations and problem formulation are introduced first. Then, we give the model of the multi-modal deep learning, followed by hashing function learning.

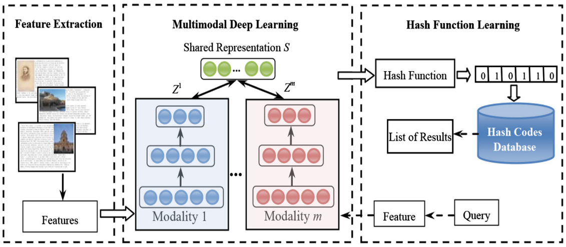

Fig. 1. Framework of the multi-modal deep learning hashing

3.1 Notations and Problem Definition

Given a set of multi-modal data $O = O^1, ..., O^p, ..., O^M (p = 1...M)$ consist of N data from M modalities, where O^p is the dataset in the p-th modal and o_i^p is the i-th datum in O^p. We use X^p to represent the features of the p-th modal, and D_p is the dimension of the feature space. Denoted the shared representation of multi-modal data is S, the projections are defined as:

$$f^p : X^p \rightarrow S^p \tag{1}$$

then, the data are mapped into hamming space using a linear projection:

$$g^p : S^p \rightarrow H^p \tag{2}$$

The main idea of learning the hash functions goes as follows. Data of each individual modal are firstly converted into the representations for single modal, denoted as B^p, which preserves the intra similarity. Data of all modals represented by B^p are then mapped into a common space S^p where the inter-similarity is preserved to generate hash functions. Finally, values of hash functions are binarized into hamming space. Given a set of multi-modal data O and the training dataset $T = (x_i^{m_i}, x_j^{m_j})^k, k = 1, ..., K$, where $x_i^{m_i} \epsilon O^{m_i}$, $x_j^{m_j} \epsilon O^{m_j}$ are the features of $o_i^{m_i}$ and $o_i^{m_j}$ separately. $L_{ij} = 1$ if two data x_j and x_j belong to the same category otherwise $L_{ij} = -1$. The distance of the two data in the shared representation is defined as:

$$d(x_i^{m_i}, x_j^{m_j}) = \|s_i^{m_i} - s_j^{m_j}\|_F^2 \tag{3}$$

We formulate the problem to the following optimal problem with the object function:

$$\min_{f} \sum_{k=1}^{K} L_{ij} d(x_i^{m_i}, x_j^{m_j}). \tag{4}$$

3.2 Multi-modal Deep Learning

In this section, we describe the multi-modal feature learning model for the task of shared representation learning, where the inputs are the features of each modal. The multi-modal deep learning consists of two components: (1) feature learning for each single modal; (2) shared feature learning for multi-modal features. Figure 2 is the deep neural network structure for the multi-modal deep learning. The whole model is learned in three steps: First, the unlabeled data U of each modal is used to pre-training the deep learning network using SDA (seeing 3.2.1). Then, the multi-modal data O is represented using the SDA of each modal and the outputs are inputted into RMB to learn the relationship between multiple modals. Finally, the training data T is used to update the parameters of the model.

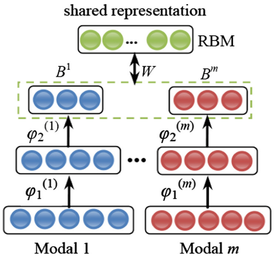

Fig. 2. Multi-modal deep learning model

Feature Learning for Single Modal. The SDA (Stacked Denoising Autoencoder [19]) is adopted to pre-training the network, which adds noises into training data based on autoencoder. Figure 3 is the process of a SDA. First, we construct the noisy version of x through a stochastic mapping. Then the noisy version x' is mapped through AE to a hidden representation $y = \varphi(x')$, where y is used to reconstruct a clean version of x by $z = \psi(y)$. Several DAs are stacked to build a layer structure, where the output of the bottom layer is the input to the higher layer. Once the encoding function is learned, encoding function is not needed anymore. We use a non-linear one-layer neural network as the unit of SDA, where the encode function is:

$$y = \varphi(x) = sigmoid(Qx + r) \tag{5}$$

and decoding function is:

$$z = \psi(y) = sigmoid(Q'y + t) \tag{6}$$

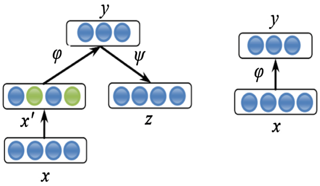

Fig. 3. Denoising Autoencoder

Multi-modal Feature Learning. After learned the representation of each modality, we use RBM (Restricted Boltz-mann Machine [15]) to model the relations between different modals and learn the shared representation of them. A Restricted Boltzmann Machine is an undirected graphical modal with stochastic visible unit v and stochastic hidden unit h. Each visible unit connects to each hidden unit, but no connections within hidden variables or visible variables. The structure of the model is shown in Fig. 4. The model defines the following energy function E:

$$E(v, h; \theta) = -a^T v - b^T h - v^T W h \tag{7}$$

where $\theta = \{a, b, W\}$ are the model parameters. The joint distribution over the visible and hidden units is defined by:

$$p(v, h; \theta) = \frac{1}{Z(\theta)} exp(-E(v, h; \theta)) \tag{8}$$

where $Z(\theta)$ is a constant for normalization. The j-th hidden node is set to 1 with probability:

$$p(h_j|v) = sigmoid(\frac{1}{\sigma^2}(b_j + W_j^T v)) \tag{9}$$

We minimize the loss function between reconstructed data using the model and original data, and learn the parameter following [7]. After obtaining the shared representation s by the multi-modal deep learning model, we can compute the derivation of the objection function with respect to $s_i^{m_i}$ and $s_j^{m_j}$ as follows:

$$\frac{\partial J}{\partial s_i^{m_i}} = 2 \sum_{k=1}^{K} L_{ij}(s_i^{m_i} - s_j^{m_j}) \tag{10}$$

$$\frac{\partial J}{\partial s_j^{m_j}} = 2 \sum_{k=1}^{K} L_{ij}(s_j^{m_j} - s_i^{m_i}) \tag{11}$$

Then, we used online gradient descent [26] to update the parameter of the last layer by:

$$W \leftarrow W - \eta \frac{J}{W} \tag{12}$$

$$b \leftarrow b - \eta \frac{J}{b} \tag{13}$$

where the derivative are computed as follows:

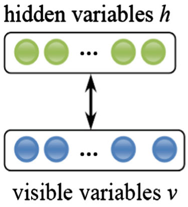

Fig. 4. Restricted Boltzmann Machine

$$\frac{\partial J}{\partial W} = \frac{\partial J}{\partial s_i^{m_i}} \frac{\partial s_i^{m_i}}{\partial W} + \frac{\partial J}{\partial s_j^{m_i}} \frac{\partial s_j^{m_i}}{\partial W} \tag{14}$$

$$\frac{\partial J}{\partial b} = \frac{\partial J}{\partial s_i^{m_i}} \frac{\partial s_i^{m_i}}{\partial b} + \frac{\partial J}{\partial s_j^{m_i}} \frac{\partial s_j^{m_i}}{\partial b_j^m} \tag{15}$$

Finally, we adopt back propagation [14] to update the parameter in the other layers of the network.

3.3 Hashing Function Learning

Let us denote the shared representation for a data is s, the linear transformation to hash code is:

$$g(s) = sign(P^T s) \tag{16}$$

where P is the projection matrix, s is the shared representation of data. Denote $S = [S^1, ..., S^M]$ as the representation for all dataset. Since our representation S is sparse, we follow the method in [22] to learn the projection matrix P by:

$$P = \sqrt{MN} \Lambda^{-\frac{1}{2}} V \Sigma^{-\frac{1}{2}} \tag{17}$$

Algorithm 1. Multi-modal deep learning based cross modal hashing

```
Input:multi-modal data $O$,training data $U$, $T$
Output:projection $f^p$,projection $g^p$

for $m = 1 : M$
    pretrainning the SDA for modality $m$
end
pretrainning the RBM
do
    for $k = 1 : K$
        $(x_i, x_j) \leftarrow T$
        update the parameter $W, b$
        update the parameter in the lower layer using back propagation
    end
Untile object function convergence
Compute $P$ using equation(17)
```

where M and N are the number of modal and the multi-modal data, V and Σ are the c largest eigenvalue and corresponding eigenvector of the matrix $\Lambda^{-\frac{1}{2}}SS^T\Lambda^{-\frac{1}{2}}$ with $\Lambda = diag(S)$. Algorithm 1 summarizes the multi-modal deep learning based cross modal hashing. Given a new data, the hash code is generated by two steps: First, extract the feature of the data and use the multi-modal deep learning to represent the data. Then, use the linear project function g to compute the hash code of the data.

4 Experiments

We evaluate our method on two real-world datasets for cross modal similarity search and analyse the results. In detail, the datasets consist of text and images, and we use text as query to search similar images and image as query to search similar texts. First, we introduce the dataset and the setting of the experiments. Then we will show the results and compare the results with other methods.

4.1 Data Sets and Settings

Two datasets are used in our experiment: Wikipedia-Picture of the Day and NUS-WIDE. All of them include two modals (pictures and text). Wikipedia [13] includes 2866 multimedia documents collected from Wikipedia website, in which each document includes one picture and at least 70 words. The dataset provides the topic probability of each text on 10 categories (computed using LDA [2]). Existing experiments used the topic probability as text features, which is too sparse to be a suitable input to deep learning. So we extract the vector space modal of each text as the feature. The feature of images use SIFT descriptor [11] based on bag-of-visual word model, which quantizes the descriptors into 1,000 dimensional vectors. The NUS-WIDE dataset is a real-world image dataset collected by Lab for Media Search in National University of Singapore [4]. It includes 81 categories and 269,648 images. Each image corresponds to multiple tags, and each image-text pair is annotated by at least one category. The image is represented by 1000-dimensional bag-of-visual word of SIFT descriptors. And the text corresponding to the image is represented by a 1000-dimensional vector of tags.

4.2 Evaluation Metric

We use mean Average Precision [23] as the evaluation metric for effectiveness in our experiment. The evaluation metric has been widely used in literatures [23,24]. The mAP evaluates the performance of similarity search, which the larger value indicates better performance and the similar results have high ranks. Given a query and R retrieved instances, the average precision is defined as:

$$AP = \frac{1}{L}\sum_{r=1}^{R} P(r)\delta(r) \tag{18}$$

where L is the number of relevance instances in the result. $P(r)$ is the accuracy of top r instances. $\delta(r)$ is indicator function, which equals to 1 if the r-th instance is relevant to the query or 0 otherwise. The mAP is the mean of all the APs and we set $R = 100$ in our experiments.

4.3 Compared Methods

We compared our method with other four cross modal hash methods. They are CMSSH, CVH, LSSH and IMVH. CMSSH [3] constructed two groups of linear hash function to keep similarity relationship between different modalities. CVH [10] extended uni-modality spectral hashing to multi-modal, and kept the similarity relationship between different modal and in the same modal. LSSH [24] adopted matrix factorization and sparse coding to map text and image into the latent factor space separately. IMVH [8] kept the interior and exterior similarity, and added the distinctive into data belongs to different categories. CMSSH and CVH generate different hash codes for different modals, but they make sure that the hash codes in the same modal have the same length.

4.4 Results

Results on Wiki Dataset. We select 90 % of the dataset as training data, 5 % as unlabeled data and the rest as the query set for MDLH. Other methods use the 95 % of the dataset (training data and unlabeled data for MDLH) as training data and the rest as the query set. The mAP of our method and compared methods on Wiki dataset are shown as Table 1. We can observe that MDLH outperforms most of the methods on two cross modal similarity search tasks. The results of existing work reported better performance on task ‘Text query Image’ than task ‘Image query Text’, because they used topics rather than words as the text feature so the text queries are represented as the 10 topic, which simplify the research problem. Furthermore, we report the Top-N precision curve of results on Wiki dataset in Fig. 5, which reflects the change of precision with respect to the number of retrieved instances.

Results on NUSE-WIDE Dataset. Some categories in NUSE-WIDE are scarce, so we select 8 categories that contain more instances than the other categories. We select 90 % of the dataset as the training data, 5 % as unlabeled data, and 5 % as the query data for MDLH. The mAP of all the methods on NUW-WIDE dataset is shown in Table 2. The performance of all methods increased to some degree on NUS-WIDE dataset.

Results on Noised Dataset. To evaluate the robustness to noise of each method, we add noises into Wiki and NUS-WIDE datasets separately, and compare the performance on the noise dataset. For Wiki and NUS-WIDE dataset, we select a category randomly as the source of noise separately. Some pictures and words from them are selected randomly as noise adding to the rest of the data.

Table 1. The *mAP* of different methods on Wiki dataset

Task	Method	Hash code length		
		16	32	64
Image query Text	CMSSH	0.3183	0.3275	0.2750
	CVH	0.3140	0.3345	0.2760
	LSSH	0.3730	0.3940	0.3887
	IMVH	0.3812	0.3921	0.3879
	MDLH	**0.3919**	**0.3940**	**0.4030**
Text query Image	CMSSH	0.3321	0.3173	0.3147
	CVH	0.3005	0.3322	0.3107
	LSSH	0.3552	0.3559	0.3545
	IMVH	0.3642	0.3624	0.3644
	MDLH	**0.3840**	**0.3729**	**0.3604**

Table 2. The *mAP* of different methods on NUS-WIDE dataset

Task	Method	Hash code length		
		16	32	64
Image query Text	CMSSH	0.4405	0.4389	0.3934
	CVH	0.3756	0.3729	0.3619
	LSSH	0.4517	0.4437	0.4460
	IMVH	0.4520	0.4489	0.4446
	MDLH	**0.4526**	**0.4537**	**0.4555**
Text query Image	CMSSH	0.4113	0.3984	0.3722
	CVH	0.3805	0.3629	0.3899
	LSSH	0.4271	0.4178	0.4143
	IMVH	0.4189	0.4250	0.4130
	MDLH	**0.4496**	**0.4478**	**0.4485**

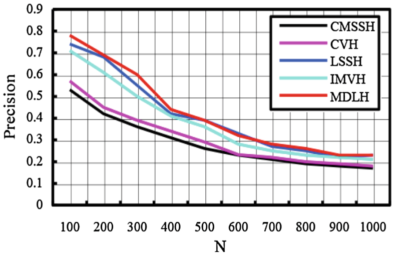

Fig. 5. Top-*N* precision of different methods on Wiki dataset

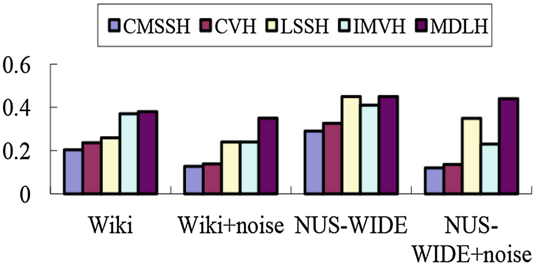

Fig. 6. The *mAP* of different methods with and without noise

In Wiki dataset, we select 2 % of the text and one picture as noise each time. In NUS-WIDE dataset, we select one tag as the noise. Figure 6 is the performance before and after adding noises. It shows that our method is robust to noise than other methods.

5 Conclusion

In this paper, we proposed a multi-modal deep learning based cross modal hash learning method. The multi-modal deep learning is used to model the relationship between multiple heterogeneous data and learn a shared representation of the multi-modal data, which is robust to noise and easy to extend to multiple modals. The experiments on two realistic dataset show that our method representing the multi-modal features effectively. In future, we will focus on the multi-modal deep learning for media types such as audio, video.

References

1. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: NIPS, pp. 153–160. MIT Press (2006)
2. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. J. Mach. Learn. Res. (JMLR) **3**, 993–1022 (2003)
3. Bronstein, M., Bronstein, A., Michel, F., Paragios, N.: Data fusion through cross-modality metric learning using similarity-sensitive hashing. In: Proceedings of the Computer Vision and Pattern Recognition, pp. 3594–3601 (2010)
4. Chua, T., Tang, J., Hong, R.: NUS-WIDE: a real-world web image database from National University of Singapore. In: CIVR (2009)
5. Datar, M., Immorlica, N., Indyk, P., Mirrokni, V.: Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of ACM Annual Symposium Computational Geometry, pp. 253–262 (2004)
6. Dahl, G., Yu, D., Deng, L., Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. TASLP **20**(1), 30–42 (2012)
7. Hinton, G., Salakhutdinov, R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
8. Hu, Y., Jin, Z., Ren, H., Cai, D., He, X.: Iterative multi-view hashing for cross media indexing. In: ACM Multimedia, pp. 527–536 (2014)

9. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1106–1114 (2012)
10. Kumar, S., Udupa, R.: Learning hash functions for cross-view similarity search. In: International Joint Conference on Artificial Intelligence, pp. 1360–1365 (2011)
11. Lowe, D.: Distinctive image features from scale-invariant key points. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
12. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML, pp. 689–696 (2011)
13. Rasiwasia, N., Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: ACM MM, pp. 251–260 (2010)
14. Rumelhart, D., Hinton, G., Williams, R.: Neurocomputing: Foundations of Research. MIT Press, Cambridge (1988)
15. Salakhutdinov, R., Hinton, G.: Deep boltzmann machines. In: AISTATS, vol. 5, pp. 448–455 (2009)
16. Sohn, K., Shang, W., Lee, H.: Improved multimodal deep learning with variation of information. In: INPS, pp. 2141–2149 (2014)
17. Song, J., Yang, Y., Yang,Y., Huang, Z., Shen, H.T.: Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: SIGMOD, pp. 785–796 (2013)
18. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep boltzmann machines. In: NIPS, pp. 2231–2239 (2012)
19. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A.: Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. JMLR **11**, 3371–3408 (2010)
20. Weiss, Y., Torralba, A., Fergus, R.: Spectral hashing. In: Advances in Neural Information Processing Systems (2005)
21. Wu, P., Hoi, S., Xia, H., Zhao, P., Wang, D., Miao, C.: Online multimodal deep similarity learning with application to image retrieval. In: ACM Multimedia, pp. 153–162 (2013)
22. Yu, Z., Wu, F., Yang, Y., Tian, Q., Luo, J., Zhuang, Y.: Discriminative coupled dictionary hashing for fast cross-media retrieval. In: SIGIR, pp. 395–404 (2014)
23. Zhen, Y., Yang, D.: A probabilistic model for multimodal hash function learning. In: SIGKDD, pp. 940–948 (2012)
24. Zhou, J., Ding, G., Guo, Y.: Latent semantic sparse hshing for cross-modal similarity search. In: SIGIR, pp. 415–424 (2014)
25. Zhu, X., Huang, Z., Shen, H., Zhao, X.: Linear cross-modal hashing for efficient multimedia search. In: ACM Multimedia, pp. 143–152 (2013)
26. Zinkevich, M.: Online convex programming and generalized infinitesimal gradientascent. In: ICML, pp. 928-936 (2003)