Comprehensive Distance-Preserving Autoencoders for Cross-Modal Retrieval

Yibing Zhan
Hangzhou Dianzi University
zybjy@hdu.edu.cn

Jun Yu*
Hangzhou Dianzi University
yujun@hdu.edu.cn

Zhou Yu
Hangzhou Dianzi University
yuz@hdu.edu.cn

Rong Zhang
University of Science and Technology of China
zrong@ustc.edu.cn

Dacheng Tao
UBTECH Sydney AI Centre, SIT, FEIT
University of Sydney
dacheng.tao@sydney.edu.au

Qi Tian
Huawei Noah's Ark Lab
University of Texas at San Antonio
qi.tian@utsa.edu
tian.qi1@huawei.com

ABSTRACT

In this paper, we propose a novel method with comprehensive distance-preserving autoencoders (CDPAE) to address the problem of unsupervised cross-modal retrieval. Previous unsupervised methods rely primarily on pairwise distances of representations extracted from cross media spaces that co-occur and belong to the same objects. However, besides pairwise distances, the CDPAE also considers heterogeneous distances of representations extracted from cross media spaces as well as homogeneous distances of representations extracted from single media spaces that belong to different objects. The CDPAE consists of four components. First, denoising autoencoders are used to retain the information from the representations and to reduce the negative influence of redundant noises. Second, a comprehensive distance-preserving common space is proposed to explore the correlations among different representations. This aims to preserve the respective distances between the representations within the common space so that they are consistent with the distances in their original media spaces. Third, a novel joint loss function is defined to simultaneously calculate the reconstruction loss of the denoising autoencoders and the correlation loss of the comprehensive distance-preserving common space. Finally, an unsupervised cross-modal similarity measurement is proposed to further improve the retrieval performance. This is carried out by calculating the marginal probability of two media objects based on a kNN classifier. The CDPAE is tested on four public datasets with two cross-modal retrieval tasks: "query images by texts" and "query texts by images". Compared with eight state-of-the-art cross-modal retrieval methods, the experimental results demonstrate that the CDPAE outperforms all the unsupervised methods and performs competitively with the supervised methods.

*Jun Yu is the corresponding author.

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240607

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; • **Computing methodologies** → *Unsupervised learning*;

KEYWORDS

Cross-Modal Retrieval; Unsupervised; Comprehensive distance-preserving; Autoencoder; Similarity Measurement

ACM Reference Format:
Yibing Zhan, Jun Yu, Zhou Yu, Rong Zhang, Dacheng Tao, and Qi Tian. 2018. Comprehensive Distance-Preserving Autoencoders for Cross-Modal Retrieval. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240607

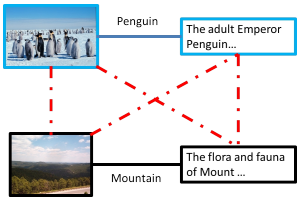
1 INTRODUCTION

The key challenge of cross-modal retrieval is how to measure the similarity between representations of different media types [11, 21]. The current mainstream methods for solving this problem are common space learning methods, which are designed to learn an intermediate common space for features of different media types and measure their similarities in the intermediate common space [11]. These methods can be grouped into three categories according to the need for label information: the supervised methods [12, 16, 19, 20, 22, 26], the semi-supervised methods [13, 27], and the unsupervised methods [1, 5–8, 23–25].

The supervised and semi-supervised methods require label information, such as the Joint Representation Learning (JRL) method [27] and the Adversarial Cross-Modal Retrieval (ACMR) method [7]. However, collecting label information is usually time-consuming and expensive in practice. The unsupervised methods rely only on cross-modal data without any additional information. Canonical Correlation Analysis (CCA) is one of the most representative works [8]. Variations of CCA include Deep Canonical Correlation Analysis (DCCA) [1] and Deep Canonically Correlated Autoencoders (DCCAE) [23]. Some unsupervised methods directly transform features of different media types to the common space without using CCA, such as the Correspondence Autoencoders (Corr-AE) [6].

Even though the previous unsupervised cross-modal retrieval methods perform well, there are still two problems to solve: 1) how to reduce the negative influence of redundant noises in features, and 2) how to directly use the relationships between representations of different objects. To the best of our knowledge, there is relatively

(a) The problem of redundant noises: the SIFT features in the backgound may affect the retrieval for the cat image.



(b) The problem of relationships between features from different objects: the relationships represented by the dotted lines are not considered in most unsupervised methods.

Figure 1: Examples of the two problems in unsupervised cross-modal retrieval.

little research regarding these two problems related to unsupervised cross-modal retrieval. Illustration of the two problem is provided in Figure 1. First, the features used in unsupervised cross-modal retrieval may contain redundant noises without the assistance of label information. Figure 1 (a) provides a query image of a cat and its Scale Invariant Feature Transform (SIFT) features. Without the help of label information to select proper features, the SIFT features from the whole image are used for retrieval. The features from the background, which are regarded as redundant noises, are also included and may influence the retrieval performance. Therefore, a strategy is needed to avoid potential negative effects caused by these noises. Second, most of the previous methods only consider the corresponding pairwise relations. The relationships among features in different objects are not directly considered and expected to be implicitly preserved by their nearby objects. Figure 1 (b) illustrates several relationships amid four representations extracted from two modalities of two objects. The solid lines indicate pairwise relations, meaning that the corresponding representations co-occur and belong to the same object. Such pairwise relations have been used in most unsupervised cross-modal retrieval methods. However, only considering these pairwise relationships may fail to account for the similarities among features extracted from different objects, which are represented by the dotted lines. In order to improve the retrieval performance [10, 22], the relationships among features from different objects should also be explicitly considered.

In this paper, we propose a novel method called unsupervised cross-modal retrieval with comprehensive distance preserving autoencoders (CDPAE) to address the two problems mentioned above. To alleviate the noise problem, we use denoising autoencoders. We notice that the procedure, which sets part of the input elements as zeroes, simulates the removal of the redundant noises from the inputs. Therefore, it reduces the negative influence of noises shown in Figure 1 (a). To explicitly calculate the similarity between features from different modalities, CDPAE proposes a comprehensive distance-preserving common space to consider all of the relationships between two representations, no matter whether they belong to the same object or not. The preserved distances include pairwise distances, heterogeneous distances, and homogeneous distances. Pairwise distances reflect the relationships between two representations of the same object from cross media types. Ideally, the pairwise distance equals zero. Heterogeneous and homogeneous distances represent the relationships between two representations extracted from two objects in cross modalities and in the same modalities, respectively. Since distances between features in the same modality generally reflect their similarities, we require the heterogeneous and homogeneous distances between two features of two objects to be equal to the distances in their original media spaces. The CDPAE incorporates the reconstruction and the distance measures into a single process by using a joint loss function. Moreover, inspired by JRL [27], a novel unsupervised similarity measurement is proposed to further improve the retrieval performance. This improvement is accomplished by calculating the the cross-modal similarity using the marginal probability of two representations based on a kNN classifier. The CDPAE is tested on four datasets and compared with eight common space learning methods. The experimental results demonstrate that the CDPAE achieves 12.5% improvement from the unsupervised methods on average and performs competitively with the semi-supervised and supervised methods.

Our contributions can be summarized as: 1) We use denoising autoencoders to reduce the negative effects of redundant noises. We propose a comprehensive distance-preserving common space to explore the relationships between representations from different modalities. As far as we know, this is one of the first works that these problems are directly discussed in unsupervised cross-modal retrieval. 2) We propose a novel unsupervised similarity measurement to calculate the similarity between two representations in the common space using marginal probabilities. Moreover, this measurement can be applied to other unsupervised methods to further improve their performance. 3) A novel and effective method, the CDPAE, is proposed for unsupervised cross-modal retrieval. The CDPAE incorporates the denosing autoencoders, the comprehensive distance-preserving common space, and the unsupervised similarity measurement. When compared with other state-of-the-art methods, the experimental results reveal that the CDPAE performs effectively and competitively. The rest of this paper is organized as follows. Section 2 gives a brief introduction of unsupervised cross-modal retrieval methods. Section 3 explains the details of the CDPAE. The experimental results and the discussions are provided in Section 4, and the conclusions are presented in Section 5.

2 RELATED WORK

Unsupervised cross-modal retrieval can be divided into two categories: 1) binary representation learning or cross-modal hashing [9, 17] and 2) real-valued representation learning. This section briefly introduces the real-valued methods.

Common space learning methods are the most popular methods for unsupervised cross-modal retrieval. Their crux is to transform features from different media types into a common space. CCA [8] learns a common space that maximizes the pairwise correlations between two sets of heterogeneous data. DCCA [1] combines the Deep Neural Network (DNN) with CCA to learn a non-linear transformed common space. DCCAE [23] improves DCCA by adding reconstruction errors based on autoencoders. Corr-AE [6] proposes the correspondence autoencoders, which contain two restricted Boltzmann machines at the code layers and jointly considers the reconstruction errors and the correlation losses using autoencoders. Similar to Corr-AE, Multi-Modal Stacked Autoencoders (MSAE)

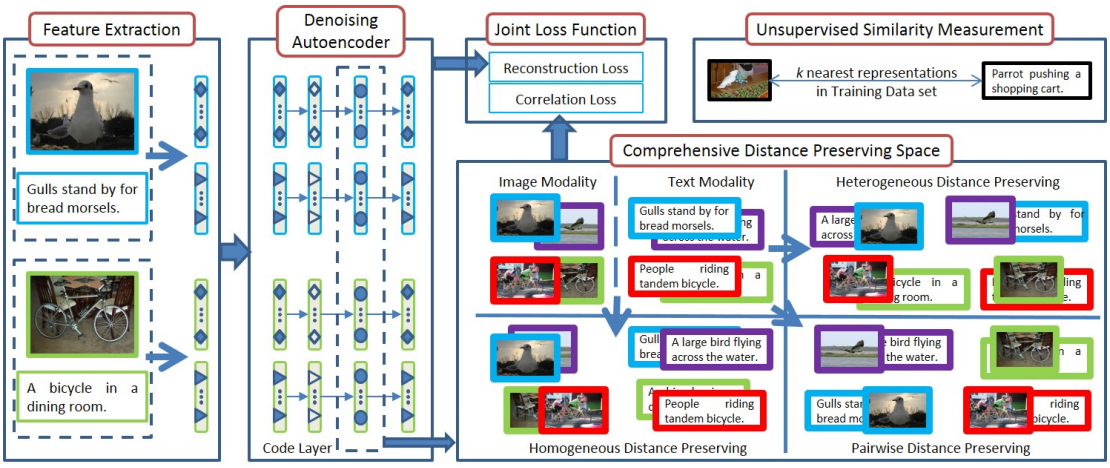


Figure 2: The overall framework of the proposed CDPAE. The CDPAE contains four parallel denoising autoencoders. The comprehensive distance-preserving common space is defined on the code layer, in which three kinds of distances are preserved according to the input representations. We use a joint loss function to combine the reconstruction loss and the correlation loss. Finally, an unsupervised cross-modal similarity measurement is applied for better retrieval performance.

[24] also consider the reconstruction errors and correlation losses based on stacked autoencoders. Deep Correspondence Restricted Boltzmann Machines (Corr-RBM) [5] use deep RBMs to preserve the pairwise relationships between different modalities. UCLA [7] uses adversarial learning to generate modality-invariant transforms.

Other types of unsupervised methods include topic model methods and cross-media similarity measurement methods [21]. Topic model methods use topics as the shared latent variables, such as Correspondence Latent Dirichlet Allocation (LDA) [2] and Topic-Regression Multi-modal Latent Dirichlet Allocation [14]. Cross-media similarity measurement methods are proposed to directly compute the cross-media similarities without obtaining an explicit common space, such as Clinchant *et al.* [4].

In summary, unsupervised cross-modal retrieval methods perform well when applied to flexible conditions. However, most previous methods rely only on pairwise relationships and have no strategies for preventing noise. It is still a challenging task to fully utilize the relationships amongst data from different media types without label information.

3 COMPREHENSIVE DISTANCE-PRESERVING AUTOENCODERS

The Comprehensive Distance-Preserving Autoencoders (CDPAE) consist of four parts: denoising autoencoders, a comprehensive distance preserving common space, a joint loss function, and an unsupervised cross-modal similarity measurement. In this section, we first describe the overall framework of the proposed CDPAE. Then, the four parts are explained in detail. Without a loss of generality, image and text modalities are used for illustration.

3.1 Overall Framework

As shown in Figure 2, the CDPAE consists of four parallel denoising autoencoders. Two of the denoising autoencoders are related to image features, and the rest are related to text features. The denoising autoencoders, which are responsible for the same modality, share the same parameters, so representations of the same modalities also have the same transforms. In each iteration, four representations between two modalities extracted from two objects are used as inputs. In this way, the relationships of different representations can be calculated on the code layer, no matter whether these representations belong to the same object or same modality. The comprehensive distance-preserving common space is defined on the code layer by preserving three kinds of distances among the input representations. The preserved distances include pairwise distances, heterogeneous distances, and homogeneous distances. All the distances are measured according to the distances between corresponding objects in the image and text modalities. Afterwards, a novel joint loss function is used to simultaneously reduce the reconstruction loss of denoising autoencoders and the correlation loss of the comprehensive distance-preserving common space. Finally, an unsupervised cross-modal similarity measurement is proposed to further improve the retrieval performance. This is carried out by using a kNN classifier on the closest training examples.

The notations used in CDPAE are defined as follows. We suppose that $V = \{v\}$ is a set of images, and $T = \{t\}$ is the corresponding set of texts. As shown in Figure 2, we let two groups of corresponding features from two objects, which are (v_i, t_i) and (v_j, t_j), represent the inputs of the CDPAE. The zeroing process $Z(*)$, encoding process $F(*)$, and decoding process $G(*)$ are defined as:

$$y = Z(x, \alpha), \quad y = F(x, \omega), \quad y = G(x, \theta). \tag{1}$$

Here, x and y represent the input and the output. α represents the proportion of the component number of x that are set as zero to the whole component number of x. ω and θ are the parameters used in $F(*)$ and $G(*)$, respectively. We let α_V, ω_V, and θ_V represent the parameters of the image modality. α_T, ω_T, and θ_T represent the parameters of the text modality.

3.2 Denoising Autoencoder

In the denoising autoencoder, a fixed number of input components are randomly set to zero, while the others are left untouched [18]. This procedure simulates the removal of the redundant noises from the inputs; therefore, it reduces the negative influence of such noises. In addition, the zeroing process can be viewed as a process of data augmentation, and it strengthens the connections between the local structures within the representation that were drawn from different modalities.

The reconstruction loss of denoising autoencoders is defined as:

$$\begin{aligned} L_{recon} =& L_r(v_i; \alpha_V, \omega_V, \theta_V) + L_r(v_j; \alpha_V, \omega_V, \theta_V) \\ &+ L_r(t_i; \alpha_T, \omega_T, \theta_T) + L_r(t_j; \alpha_T, \omega_T, \theta_T), \end{aligned} \tag{2}$$

where

$$L_r(x; \alpha, \omega, \theta) = \|x - G(F(Z(x, \alpha), \omega), \theta)\|_2 . \tag{3}$$

Here, $\|*\|_2$ is the L^2 norm. L_r represents the reconstruction loss caused by a single input x.

3.3 Comprehensive Distance-Preserving Common Space

The CDPAE uses the cosine distance to measure the similarity of features in the same media spaces. The distance is calculated as:

$$C(X, Y) = \frac{1 - X \cdot Y}{\|X\| \, \|Y\|} = 1 - \sum_{k=1}^{n} x_k y_k / \sqrt{\sum_{k=1}^{n} x_k^2} \sqrt{\sum_{k=1}^{n} y_k^2}. \tag{4}$$

Here, X and Y are two vectors with the same size of n. x_k and y_k are the kth components of X and Y, respectively. The cosine distance is selected because it is commonly used in cross-modal retrieval. According to Eq. (4), the distance between two inputs for the CDPAE is measured as:

$$D(x_1, x_2) = C(F(Z(x_1, \alpha_1), \omega_1), F(Z(x_2, \alpha_2), \omega_2)). \tag{5}$$

Here, x_1 and x_2 stand for any two representations, no matter the respective media types. α_i and ω_i are the parameters of the corresponding media type of x_i, $i = 1, 2$.

As shown in Figure 2, there are three kinds of distances in the comprehensive distance-preserving common space: pairwise distances, heterogeneous distances, and homogeneous distances.

Pairwise distances reflect the relationships between representations of the same objects in cross modalities, and they are expected to be as small as possible. The pairwise distance loss is defined as:

$$L_{pair} = D(v_i, t_i) + D(v_j, t_j). \tag{6}$$

Preserving pairwise distances leads to a common space, in which representations belonging to the same objects cluster together.

Heterogeneous distances reflect the relationships between representations of different objects in cross modalities. We limit them to be consistent with the distances between the corresponding objects in their original media spaces. The heterogeneous distance loss is defined as:

$$L_{heter} = |D(v_i, t_j) - d| + |D(v_j, t_i) - d|, \tag{7}$$

where

$$d = \sqrt{C(v_i, v_j) C(t_i, t_j)}. \tag{8}$$

Preserving heterogeneous distances results in a common space, in which representations of different objects from cross modalities get close, only if the distances between these representations are small in both the original image and text spaces. If distances between these representations in the original media spaces are large, then the representations do not cluster together.

Homogeneous distances measure the relationships between representations from different objects in the same modalities. Because, in each iteration, heterogeneous and homogeneous distances are calculated between the same two objects, their values are set as the same in the CDPAE. The homogeneous distance loss is defined as:

$$L_{homo} = |D(v_i, v_j) - d| + |D(t_i, t_j) - d|. \tag{9}$$

Preserving homogeneous distances lead to a common space, in which representations of different objects in the same modalities get close, only if the distances between these representations are small in both the original image and text spaces.

The comprehensive distance-preserving common space combines the three distance measures mentioned above. Consequently, it contains all the advantages possessed by all three of the common spaces obtained by single distance-preserving measures. Because heterogeneous and homogeneous distances are extracted from the same two objects, we believe that the heterogeneous and homogeneous distance loss have the same significance. Thus, the correlation loss in the comprehensive distance-preserving common space is defined as follows.

$$L_{corr} = L_{pair} + \lambda_1 (L_{heter} + L_{homo}). \tag{10}$$

Here, λ_1 is the parameter to trade off between the pairwise distance loss and other distance losses.

3.4 Joint Loss Function

A joint loss function is proposed to simultaneously calculate the reconstruction loss of the denoising autoencoders and the correlation loss of the comprehensive distance preserving common space [6]. The joint loss function is defined as follows.

$$L = L_{corr} + \lambda_2 L_{recon}. \tag{11}$$

Here, λ_2 is the parameter used to trade off between two groups of objectives: the reconstruction loss and the correlation loss. By combining Eq. (10) and Eq. (11), the joint loss function is given as:

$$L = L_{pair} + \lambda_1 (L_{heter} + L_{homo}) + \lambda_2 L_{recon}. \tag{12}$$

3.5 Unsupervised Cross-Modal Similarity Measurement

So far, we have described how to learn a common space using the CDPAE. Next, an unsupervised cross-modal similarity measurement is proposed to further improve the retrieval performance.

The retrieval distances between transformed features within the training dataset in the common space usually have more confidence than the distances within the testing data. Therefore, the similarity between two features is defined as the marginal probability based on a kNN classifier that is used to classify each media representation in the training examples. The similarity of two representations is defined as:

$$Sim(v, t) = \sum_{i=1}^{k} \sum_{j=1}^{k} P(l_{p_i} = l_{q_j}) \cdot P(l_v = l_{p_i} | v, p_i) \cdot P(l_t = l_{q_j} | t, q_j), \tag{13}$$

where l_x represents the category of x. v and t denote any two representations from images and texts. $\{p_i\}_{i=1}^k$ and $\{q_j\}_{j=1}^k$ are the top k nearest representations in the training data sets of v and t, irrespective of the media types.

We suppose that the distance between two representations reflects the possibility that they belong to the same semantic category. Thus, if two representations in the training data set correspond pairwise, their possibility is set as 1; otherwise, the possibility is measured as:

$$P(l_{p_i} = l_{q_j}) = 1 - D(p_i, q_j)/2. \tag{14}$$

By using Eq. (14), the range of possibilities for the cosine distance is between 0 and 1. A lower distance means that there is a higher possibility that the corresponding representations belong to the same category. However, better calculations for these possibilities may exist. In the same manner, the conditional possibility that one testing representation and its k nearest training data belong to the same category is defined as:

$$P(l_v = l_{p_i}|v, p_i) = \frac{1 - D(v, p_i)/2}{\sum_{i=1}^k (1 - D(v, p_i)/2)}, \tag{15}$$

$$P(l_t = l_{q_j}|t, q_j) = \frac{1 - D(t, q_j)/2}{\sum_{j=1}^k (1 - D(t, q_j)/2)}. \tag{16}$$

4 EXPERIMENTS

4.1 Datasets

Four public real-world datasets are used for algorithm validation and comparison: Wikipedia [16], NUS-WIDE-10K [3], Pascal Sentence [15], and XMedia [13, 27].

The Wikipedia dataset [16] consists of 2,866 image/text pairs from ten categories. The dataset is split into three subsets. There are 2,173 cases in the training set, 231 cases in the validation set, and 462 cases in the testing set. Two kinds of representations are used in the experiments. In the Wikipedia-Multiple[1] dataset, image features contain 2,296 dimensions, including 1,000 dimensional Pyramid Histograms of Words, 512 dimensional GISTs, and 784 dimensional MPEG-7 descriptors. The text features contain 3,000 dimensional bag-of-words (BOW) vectors. In the Wikipedia-Shallow[2] dataset, image features contain a 128 dimensional BOW histograms from a SIFT code book with 128 code words. The text features include 10 dimensional histograms from a 10-topic LDA model.

The NUS-WIDE-10K dataset [3] has 10,000 image/text pairs selected evenly from the ten largest categories in the NUS-WIDE dataset. The dataset is split into three subsets. There are 8,000 pairs in the training set, 1,000 pairs in the validation set, and 1,000 in the testing set. In the NUS-WIDE-10K[1] dataset, image features have 1,134 dimensions, which are the concatenation of 64 dimensional color histograms, 144 dimensional color correlograms, 73 dimensional edge direction histograms, 128 dimensional wavelet textures, 225 dimensional block-wise color moments, and 500 dimensional SIFT-based BOW features. The text features include 1,000 dimentional BOW vectors.

The Pascal Sentence dataset [15] contains 1,000 image/text pairs, which are evenly categorized into 20 categories. The dataset is split into three subsets. There is a training set with 800 pairs (40 cases per category), a testing set with 100 pairs (5 cases per category), and a validation set with 100 pairs (5 cases per category). In the Pascal Sentence[1] dataset, the image representations contain 2,296 dimensions, including 1,000 dimensional Pyramid Histograms of Words, 512 dimensional GISTs, and 784 dimensional MPEG-7 descriptors. The text representations are 1,000 dimensional BOW vectors.

The XMedia dataset [13, 27] contains 5,000 image/text pairs with 20 categories. The dataset is randomly split into three subsets. There are 4,000 cases in the training set, 300 cases in the validation set, and 700 cases in the testing set. Two kinds of representations from the XMedia dataset are used to do the experiments. For the XMedia-Deep[3] dataset, image features include 4,096 dimensional CNN features extracted by the fc7 layer of AlexNet. The text features constitute 3,000-dimensional BOW vectors. In the XMedia-Shallow[3] dataset, image features consist of 128 dimensional BOW histograms from a SIFT code book with 128 code words. The text features contain 10 dimensional histograms from a 10-topic LDA model.

As we can see, these datasets have very distinct properties. The sizes of these datasets range from 1,000 to 10,000 units, and the number of categories ranges from 10 to 20. For each modality, there are multiple kinds of representations.

4.2 Baselines and Evaluation Metric

We compare the CDPAE with eight state-of-the-art cross-modal retrieval methods: CCA [8], MSAE [24], DCCAE [23], Corr-AE [6], CMCP [26], JRL [27], JFSSL [20], and ACMR [19]. CCA, MSAE, DCCAE, and Corr-AE are unsupervised methods. JRL is a semi-supervised method. CMCP, JFSSL, and ACMR are supervised methods, respectively.

Two cross-modal retrieval tasks are used for testing: text retrieval from an image query (Img2Txt) and image retrieval from a text query (Txt2Img). The retrieval performance is evaluated using Mean Average Precision (MAP). Given one query and the first R top-ranked retrieved data, the average precision (AP) is defined as:

$$AP = \frac{1}{M} \sum_{k=1}^{R} \frac{M_k}{k} \cdot rel_k. \tag{17}$$

Here, M is the number of relevant data in the retrieved results, M_k is the number of relevant items in the top k returns, and rel_k, represents the relevance of a given rank. $rel_k = 1$ if the item ranked at the kth position is relevant; otherwise, it is zero. The MAP is obtained by averaging the AP of all the queries. We report the MAP@50 ($R = 50$) in all experiments following [6, 19]. In addition, we also display the precision-scope curves for all of the methods.

4.3 Implementation Details

In the CDPAE, for all of the datasets, the size of the code layer is set to 1,024. The batch training size is set to 64. The *tanh* function is used to nonlinearly project raw features into the common subspace following [19]. The Adam Optimizer with the learning rate set to 0.0001 is used to train the CDPAE. All the weights have L_2 regulations with a weighted coefficient set to 0.5.

Five parameters, α_V, α_T, λ_1, λ_2, and k, need to be tuned for the CDPAE. For the sake of simplicity and time-saving, we perform a

[1] https://github.com/FangxiangFeng/deepnet/tree/master/deepnet/examples/mm14
[2] http://svcl.ucsd.edu/projects/crossmodal/
[3] http://www.icst.pku.edu.cn/mipl/XMedia/

Table 1: Performance comparison of cross-modal retrieval methods on four datasets. * denotes supervised methods. Δ denotes semi-supervised method. The rest are unsupervised methods.

Tasks		CCA	MSNE	DCCAE	Corr-AE	CMCP*	JRL$^{\Delta}$	JFSSL*	ACMR*	**CDPAE**
Wikipedia-Multiple	Img2Txt	0.175	0.291	0.269	0.336	**0.370**	**0.373**	0.358	0.346	**0.393**
	Txt2Img	0.176	0.326	0.280	0.361	**0.503**	0.437	0.439	**0.491**	**0.466**
	Avg.	0.176	0.309	0.275	0.349	**0.437**	0.405	0.399	**0**.419	**0.429**
Wikipedia-Shallow	Img2Txt	0.255	0.241	0.259	0.272	**0.306**	**0.308**	0.273	0.272	**0.277**
	Txt2Img	0.334	0.296	0.323	0.269	**0.388**	**0.383**	0.342	0.362	**0.366**
	Avg.	0.289	0.268	0.291	0.271	**0.347**	**0.346**	0.308	0.317	**0.321**
NUS-WIDE-10K	Img2Txt	0.291	0.306	0.343	0.331	0.392	**0.509**	**0.476**	0.347	**0.437**
	Txt2Img	0.306	0.341	0.348	0.378	0.410	**0.560**	**0.538**	0.463	**0.478**
	Avg.	0.299	0.324	0.346	0.355	0.401	**0.535**	**0.507**	0.405	**0.458**
Pascal Sentence	Img2Txt	0.126	0.246	0.253	0.281	0.261	0.295	**0.305**	**0.336**	**0.311**
	Txt2Img	0.130	0.252	0.231	0.291	0.248	0.220	**0.292**	**0.310**	**0.308**
	Avg.	0.128	0.249	0.242	0.286	0.255	0.258	**0.299**	**0.323**	**0.310**
XMedia-Deep	Img2Txt	0.122	0.799	0.772	0.855	0.831	**0.899**	**0.897**	0.895	**0.901**
	Txt2Img	0.120	0.776	0.773	0.882	0.765	**0.934**	0.930	**0.931**	**0.947**
	Avg.	0.121	0.788	0.773	0.869	0.798	**0.917**	**0.914**	0.913	**0.924**
XMedia-Shallow	Img2Txt	0.199	0.179	0.215	0.119	**0.258**	**0.267**	0.219	**0.280**	0.215
	Txt2Img	0.196	0.173	0.222	0.113	**0.259**	**0.248**	0.201	**0.257**	0.215
	Avg.	0.198	0.176	0.219	0.116	**0.259**	**0.258**	0.210	**0.269**	0.215

3-step grid search for the parameters. First, the values of α_V and α_T are fixed at 0.5, and we use the grid search to tune λ_1 from 0.1 to 1.9 at an increment of 0.2 per step and λ_2 from 0.001 to 1 ten times per step. Then, the values of λ_1 and λ_2 are fixed according to the best performing values in the first step. We tune both α_V and α_T using the grid search at an increment of 0.1 per step from 0 to 0.9. Finally, the unsupervised cross-modal measurement is added, and we tune k using the grid search at an increment of 10 per step from 10 to the size of the retrieval dataset. Only the validation sets are used to determine the parameters. The parameters of the other methods are also adjusted according to the validation sets. During the data preprocessing stage, the CDPAE standardizes all representations, except for the BOW vectors, which require no preprocessing step. Data preprocessing is also used in other methods to improves their performance.

4.4 Performance Comparisons

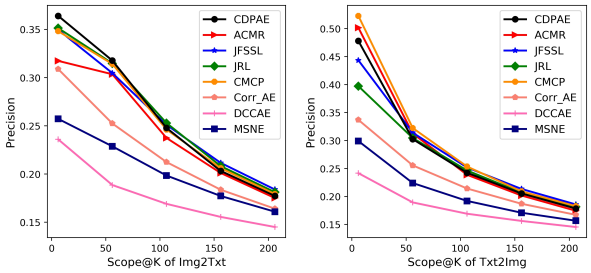
In this subsection, the performance of the CDPAE is compared with eight state-of-the-art cross-modal retrieval methods. The MAP@50 scores for all the cross-modal retrieval tasks are provided in Table 1. For each performance measurement, the top three methods are highlighted in boldface.

1) The CDPAE achieves significant improvements in nearly all the retrieval tasks when compared with the unsupervised methods, CCA, MSNE, DCCAE, and Corr-AE. In particular, in Wikipedia-Multiple, Wikipedia-Shallow, NUS-WIDE-10K, Pascal Sentence, and XMedia-Deep, the proposed CDPAE outperforms the best unsupervised competitor by 22.9%, 10.3%, 29.0%, 8.4%, and 6.3% on average, respectively. In the case of the XMedia-Shallow dataset, the CDPAE is the second best performer, only 1.8% lower than the best method, DCCAE; nevertheless, the CDPAE is still 8.6% higher than the third best method, CCA. These high performances demonstrate the capacity of the CDPAE for unsupervised cross-modal retrieval. In addition, all the unsupervised methods only utilize the corresponding pairwise relationships. The relationships among different objects contributed to the final performance of the CDPAE and requires additional exploration in future research.

2) The CDPAE still performs competitively when compared with the semi-supervised and supervised methods, CMCP, JRL, JFSSL, and ACMR. It can be observed that supervised methods generally outperform unsupervised methods in all retrieval tasks. However, the CDPAE ranks in the top three in most retrieval tasks. In particular, in Img2Txt for Wikipedia-Multiple and both retrieval tasks for XMedia-Deep, the CDPAE outperforms all the other baselines. This reveals that the distance among different objects in the same modality potentially reflects their semantic correlation.

3) The CDPAE ranks in the top three for nearly all the datasets. The performance of the CDPAE is relatively robust. Other methods work well for some datasets but fail to provide satisfactory results for the rest. For example, the Corr-AE performs satisfactorily on the Pascal Sentence, but poorly on the XMedia-Shallow and Wikipedia-Shallow. The CMCP performs well on two Wikipedia datasets; however, the performance of CMCP on the XMedia-Deep dataset is unsatisfactory. The CDPAE can be applied to datasets with various features as well as different data sizes.

In addition to the evaluation measured by the MAP@50 score, we also provide precision-scope curves for comparison. Figure 3 shows the curves for all the methods, except for CCA, on the Wikipedia-Multiple dataset. It can be seen that the precision-scope evaluation is consistent with the MAP@50 scores. The CDPAE outperforms the unsupervised methods and performs competitively when compared with the supervised methods.



(a) Precision-scope curves of Img2Txt (b) Precision-scope curves of Txt2Img

Figure 3: Precision-scope curves on the Wikipedia-Multiple dataset with K ranging from 6 to 206.

4.5 Parameter Analysis

In previous experiments, we empirically set the model parameters of the CDPAE as, α_V, α_T, λ_1, λ_2, and k. Each has a special meaning to the CDPAE. In this subsection, we study the performance impacts of these individual parameter values. Two Wikipedia datasets are used as test beds. The evaluation is conducted by changing one of the observed parameters while fixing the others, as in [19]. The performances of α_V, α_T, λ_1, and λ_2 are measured by CDPAE without using the unsupervised cross-modal similarity measurement.

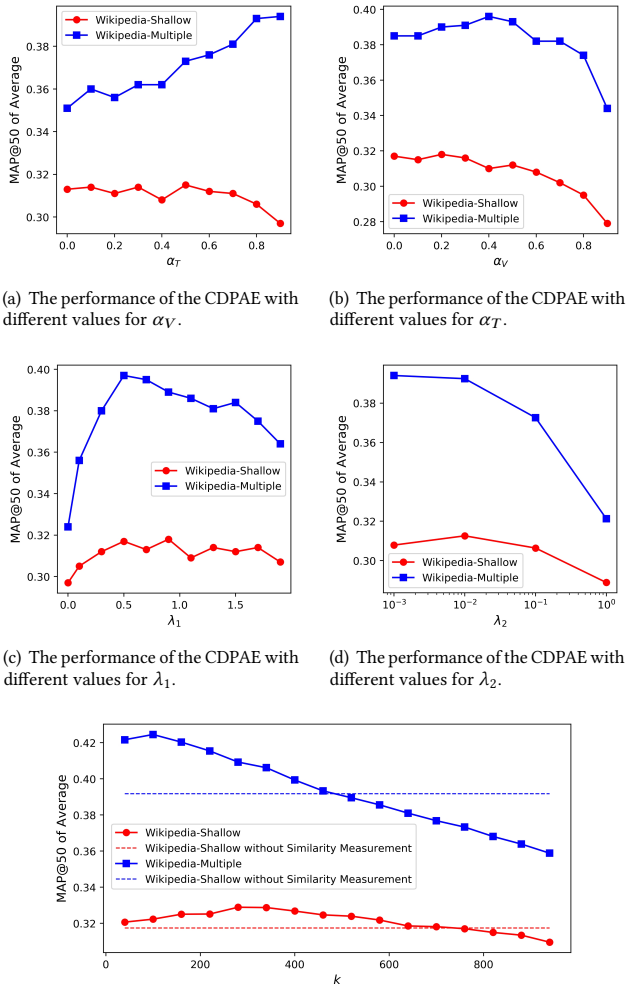
Figure 4 shows the average performance of the CDPAE with different parameter values on the two Wikipedia datasets. From Figure 4, we can draw the following conclusions.

1) α_V and α_T represent the degree of the zeroing process for the denoising autoencoders that are used for the image and text modalities, repectively. Note that $\alpha_V = 0$ and $\alpha_T = 0$ means that there is no denoising process used by the autoencoders for the corresponding modalities. From Figure 4 (a) and (b), we can observe that, for both datasets, finding proper values for α_V and α_T improves the retrieval performance. This reveals that there are redundant noises in the representations that have been extracted from different modalties, and these noises influence the final retrieval performance. The denoising autoencoder is a useful strategy and reduces the negative effects of reducent noise to some extent.

2) λ_1 determines the significance of the heterogeneous and homogeneous distance loss in the CDPAE. If $\lambda_1 = 0$, the comprehensive distance-preserving common space only preserves the pairwise distance between reprentations of the same objects in cross modalties. From Figure 4 (c), it can be seen that both higher or lower values of λ_1 result in poor performance. Considering the distances between different objects significantly improves the performance of both datasets. This again demonstrates the potential relationships between distance and semantic information concerning representations. Exploring the relationships between different objects remains a useful topic for unsupervised cross-modal retrieval.

3) λ_2 denotes the significance of the reconstruction loss in the CDPAE. It can be observed from Figure 4 (d) that reducing the reconstruction loss contributes slightly to the final retrieval performance. For both datasets, the values of λ_2 are relatively low.

4) k is the number of nearest neighbors that are required by the kNN classifier for use in the unsupervised cross-modal similarity measurement. In Figure 4 (e), we also provide the retrieval performance of the CDPAE without using the unsupervised cross-modal similarity measurement. It can observed that by using the unsupervised cross-modal similarity measurement, we can improve the results of the CDPAE in Wikipedia-Multiple and Wikipedia-Shallow datasets by 9.7% and 1.3% on average, respectively.



(a) The performance of the CDPAE with different values for α_V.

(b) The performance of the CDPAE with different values for α_T.

(c) The performance of the CDPAE with different values for λ_1.

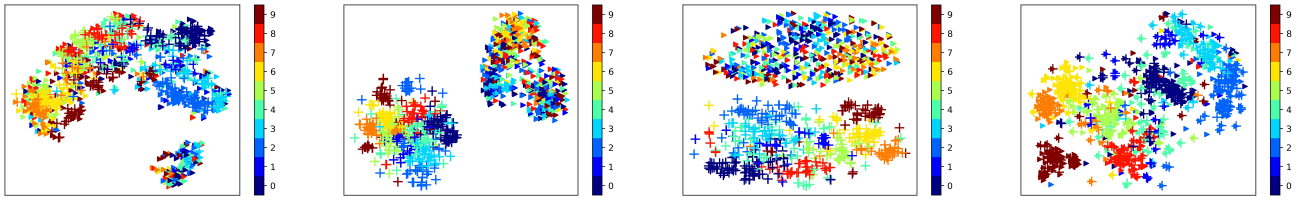
(d) The performance of the CDPAE with different values for λ_2.

(e) The performance of the CDPAE with different values for k. The dotted lines indicate the performance of the CDPAE without using the unsupervised cross-modal similarity measurement.

Figure 4: The retrieval performance of the CDPAE with different parameter values on the two Wikipedia datasets.

4.6 Visualizations of Distance-Preserving Common Spaces

In this subsection, we use the t-SNE tool, as in [19], to visualize the distribution of the transformed representations. Our trained models carry this out using 500 sample points from each modality within the Wikipedia-Multiple dataset. Transformed representations of



(a) Pairwise Distance-Preserving Common Space (b) Heterogeneous Distance-Preserving Common Space (c) Homogeneous Distance-Preserving Common Space (d) Comprehensive Distance-Preserving Common Space

Figure 5: t-SNE visualizations for the test data in the Wikipeida-Multiple dataset. Different shapes denote different modalities: "Triangle" for images and "+" for texts. Different colors denote different semantic categories.

Table 2: Performance comparison of methods with and without the unsupervised cross-modal retrieval measurement.

Methods	Wikipedia-Multiple		Wikipedia-Shallow	
	Img2Txt	Txt2Img	Img2Txt	Txt2Img
CCA	**0.175**	0.176	0.256	0.322
CCA-s	0.173	**0.183**	**0.258**	**0.339**
MSNE	**0.291**	0.326	0.241	0.296
MSNE-s	0.290	**0.344**	**0.242**	**0.301**
DCCAE	0.269	0.280	**0.259**	0.323
DCCAE-s	**0.290**	**0.302**	0.257	**0.332**
Corr-AE	0.336	0.361	0.272	0.269
Corr-AE-s	**0.350**	**0.402**	**0.279**	**0.302**

four different common spaces are provided: the pairwise distance-preserving common space, the heterogeneous distance-preserving common space, the homogeneous distance-preserving common space, and the comprehensive distance-preserving common space. Besides the comprehensive distance-preserving common space, other common spaces only consider the corresponding preserved distance loss in their final objective loss function.

Figure 5 shows the t-SNE visualization for different common spaces in Wikipedia-Multiple dataset. It can be seen that, in the pairwise distance-preserving common space, representations of images and texts tend to mix. However, the representations from the same categories are unsatisfactorily clustered. In the heterogeneous and homogeneous distance-preserving common space, representations of images and texts have obvious distinctions. This is because the intra-modality distance is much smaller than the inter-modality distance. Furthermore, in the homogenerous distance-preserving common space, representations from the same modalities cluster together according to their respective categories. In the comprehensive distance-preserving common space, the transformed representations of images and texts are distributed in the best manner. A large number of representations with the same semantic labels are clustered, irrespective of their media types. This reveals that the comprehensive distance-preserving common space contains the advantages of the other three distance-preserving common spaces, and it is quite effective for cross-modal retrieval tasks.

4.7 Applications of the Unsupervised Cross-Modal Similarity Measurement

We notice that the proposed unsupervised cross-modal similarity measurement requires almost no prerequisites. Consequently, it could be applied to any unsupervised common space learning methods. This subsection discusses the impacts of applying the unsupervised cross-modal similarity measurement to other methods.

Table 2 provides the MAP@50 scores on two Wikipedia datasets. Here, CCA-s, MSNE-s, DCCAE-s, and Corr-AE-s represent corresponding methods with the unsupervised cross-modal similarity measurement. In all the methods, k is tuned by the validation sets. The better results produced by the original method and its variation are highlighted in boldface. It can be observed that the unsupervised cross-modal similarity measurement improves the performance of all the methods in nearly all retrieval tasks. These improvements demonstrate that the proposed unsupervised cross-modal similarity measurement could be applied to different methods to improve their retrieval performance.

5 CONCLUSIONS

In this paper, a novel unsupervised cross-modal retrieval method, the Comprehensive Distance-Preserving Autoencoders (CDPAE), is proposed to solve the problem of redundant noises and to further explore the correlations among representations from different modalities. The CDPAE includes four components: denoising autoencoders, a comprehensive distance-preserving common space, a joint loss function, and an unsupervised cross-modal similarity measurement. The experimental performances of the cross-modal retrieval tasks on four public datasets demonstrate that the CDPAE outperforms all unsupervised methods by 12.5% on average. When compared with semi-supervised and unsupervised methods, the CDPAE performs competitively and achieves the top three performance on most datasets.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant No.: 61622205, 61472110, 61429201, and 61702143, in part by the Zhejiang Provincial Natural Science Foundation of China under Grant No. LR15F020002, in part by the ARC FL-170100117, DP-180103424, LP-150100671, and in part to Dr. Qi Tian by ARO under Grant No. W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar.

REFERENCES

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *International Conference on Machine Learning*. 1247–1255.

[2] David M Blei and Michael I Jordan. 2003. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 127–134.

[3] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of the ACM international conference on image and video retrieval*. ACM, 48.

[4] Stéphane Clinchant, Julien Ah-Pine, and Gabriela Csurka. 2011. Semantic combination of textual and visual information in multimedia retrieval. In *Proceedings of the 1st ACM international conference on multimedia retrieval*. ACM, 44.

[5] Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2015. Deep correspondence restricted Boltzmann machine for cross-modal retrieval. *Neurocomputing* 154 (2015), 50–60.

[6] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 7–16.

[7] Li He, Xing Xu, Huimin Lu, Yang Yang, Fumin Shen, and Heng Tao Shen. 2017. Unsupervised cross-modal retrieval through adversarial learning. In *Multimedia and Expo (ICME), 2017 IEEE International Conference on*. IEEE, 1153–1158.

[8] Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika* 28, 3/4 (1936), 321–377.

[9] Venice Erin Liong, Jiwen Lu, Yap-Peng Tan, and Jie Zhou. 2017. Cross-modal deep variational hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4077–4085.

[10] Yingwei Pan, Ting Yao, Tao Mei, Houqiang Li, Chong-Wah Ngo, and Yong Rui. 2014. Click-through-based cross-view learning for image search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 717–726.

[11] Yuxin Peng, Xin Huang, and Yunzhen Zhao. 2017. An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges. *IEEE Transactions on Circuits and Systems for Video Technology* (2017).

[12] Yuxin Peng, Jinwei Qi, Xin Huang, and Yuxin Yuan. 2018. CCL: Cross-modal Correlation Learning With Multigrained Fusion by Hierarchical Network. *IEEE Transactions on Multimedia* 20, 2 (2018), 405–420.

[13] Yuxin Peng, Xiaohua Zhai, Yunzhen Zhao, and Xin Huang. 2016. Semi-supervised cross-media feature learning with unified patch graph regularization. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 3 (2016), 583–596.

[14] Duangmanee Putthividhy, Hagai T Attias, and Srikantan S Nagarajan. 2010. Topic regression multi-modal latent dirichlet allocation for image annotation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 3408–3415.

[15] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 139–147.

[16] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 251–260.

[17] Fumin Shen, Xiang Zhou, Yang Yang, Jingkuan Song, Heng Tao Shen, and Dacheng Tao. 2016. A fast optimization method for general binary code learning. *IEEE Transactions on Image Processing* 25, 12 (2016), 5610–5621.

[18] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*. ACM, 1096–1103.

[19] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial Cross-Modal Retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 154–162.

[20] Kaiye Wang, Ran He, Liang Wang, Wei Wang, and Tieniu Tan. 2016. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE transactions on pattern analysis and machine intelligence* 38, 10 (2016), 2010–2023.

[21] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215* (2016).

[22] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5005–5013.

[23] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *International Conference on Machine Learning*. 1083–1092.

[24] Wei Wang, Beng Chin Ooi, Xiaoyan Yang, Dongxiang Zhang, and Yueting Zhuang. 2014. Effective multi-modal retrieval based on stacked auto-encoders. *Proceedings of the VLDB Endowment* 7, 8 (2014), 649–660.

[25] Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 3441–3450.

[26] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. 2012. Cross-modality correlation propagation for cross-media retrieval. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2337–2340.

[27] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. 2014. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology* 24, 6 (2014), 965–978.