Online Asymmetric Similarity Learning for Cross-Modal Retrieval

Yiling Wu[1,2], Shuhui Wang[1], Qingming Huang[1,2]
[1]Key Lab of Intell. Info. Process., Inst. of Comput. Tech, Chinese Academy of Sciences, China
[2]University of Chinese Academy of Sciences, China
yiling.wu@vipl.ict.ac.cn, wangshuhui@ict.ac.cn, qmhuang@ucas.ac.cn

Abstract

Cross-modal retrieval has attracted intensive attention in recent years. Measuring the semantic similarity between heterogeneous data objects is an essential yet challenging problem in cross-modal retrieval. In this paper, we propose an online learning method to learn the similarity function between heterogeneous modalities by preserving the relative similarity in the training data, which is modeled as a set of bi-directional hinge loss constraints on the cross-modal training triplets. The overall online similarity function learning problem is optimized by the margin based Passive-Aggressive algorithm. We further extend the approach to learn similarity function in reproducing kernel Hilbert spaces by kernelizing the approach and combining multiple kernels derived from different layers of the CNN features using the Hedging algorithm. Theoretical mistake bounds are given for our methods. Experiments conducted on real world datasets well demonstrate the effectiveness of our methods.

1. Introduction

Our world is proliferated with data of heterogeneous modalities, such as image, text, audio, etc. In real application scenarios, there is a surge of interests in generating natural sentences describing an image [20, 30, 38]. More generally, given queries in one modality, users would search for semantically relevant content in other modalities [25, 29, 21, 13], *e.g.*, images that best illustrate the topic of a textual query, or textual descriptions that best explain the content of a visual query. Therefore, effective and efficient techniques are in urgent needs to relate heterogeneous modalities and facilitate real-world cross-modal retrieval.

One of the critical problems in cross-modal retrieval task is how to measure the similarity between queries and database entries from different modalities. For example, image can be represented by a set of multi-layered feature responses of a convolutional neural network (CNN) [17], and text can be represented by bag-of-words model or an aggregation of word vectors [22]. However, the modality heterogeneity makes it a non-trivial issue that needs to be carefully considered.

As the standard solution to cross-modal correlation learning, the aim of subspace learning [25, 29, 21, 13] is to find a low dimensional latent common space that well preserve or capture the cross-modal relationship among data objects. By projecting multi-modal data into the latent space, simple distance measurement, *e.g.*, the Cosine distance [25, 29] or Euclidean distance [13], can be used to measure the semantic distance. Furthermore, semantic relation preservation is pursued directly on distance or similarity level by cross-modal metric learning [2, 10]. In this paper, we address two issues of existing works that limit their application to real-world scenarios.

First of all, existing works assume that documents of different modalities are expressed by high-dimensional vectors, although the information delivered in each modality is different. Accordingly, a shared subspace is learned by treating different modalities indiscriminately and maximizing the correlation between two modalities [25, 29, 21, 13] with a pre-specified number of dimensions of the shared subspace. However, different settings result in different model capacity and retrieval performance. To address this issue, we propose a simple but efficient asymmetric bilinear similarity measurement. The measurement is extended from existing Mahalanobis-matrix-based metric [37, 9, 6] on single modality. We relax the positive semi-definiteness property of the covariance matrix, and allow it to be determined by maximizing the cross-modal ranking performance. It can be regarded as a linear function on the joint feature of the two modalities, so their multiplicative interactions can be directly measured [2]. It can also be interpreted as inner-product of the projected representations, but we do not need any prior information on the number of latent dimensions as sub-space learning approaches. Consequently, it better facilitates real-world applications where the cross-modal content tends to be diversified.

To further address the modality heterogeneity problem, we consider the situation where documents of one modality is expressed by complex representation, *e.g.*, the CNN features on visual modality. To utilize the state-of-the-art CNN representation, the simplest way is to concatenate the multi-layered feature responses of CNN into a unified feature. However, this will lead to the curse-of-dimensionality in cross-modal correlation learning. A straightforward solution is to extend the linear metric learning functions to nonlinear settings by kernelized models, which has been extensively studied in single-modal metric learning [28, 10, 15] and cross-modal correlation learning [11, 10, 29]. The kernel metric learning has also been extended with multiple kernels [31, 34, 40]. However, the metric learning on multiple kernels has been rarely studied in cross-modal learning context.

In our study, our bilinear asymmetric similarity can be kernelized [11, 29] by using modality specific kernel functions under the structural risk minimization framework. To fully combine the multi-layered CNN representation, we propose an asymmetric multi-kernel similarity measurement that can be derived from the dual problem of structural risk minimization. The proposed asymmetric similarity, optimized by an online kernel learning procedure as [36], learns a flexible nonlinear proximity function with multiple kernels, thus the performance of cross-modal retrieval can be significantly improved by addressing the modality heterogeneity.

Second, similarities expressed in relative order [27, 21] provide a more flexible way than absolute similarity scores to represent multiple levels of relevance between heterogeneous data objects. There are two kinds of relative similarities [35], *i.e.*, the relative similarity of texts to an image and the relative similarity of images to a given text. Due to the modality heterogeneity, the relative similarity in one direction can not be used to infer the relative similarity in another. Both of them are equally important for learning similarity function to combine the domain-specific properties in both visual and textual modalities. To optimize the retrieval performance, we propose to learn the asymmetric linear and multi-kernel similarities from two relative similarity directions to preserve the relative order for data with single label and multiple labels. Consequently, the cross-modal retrieval performance can be directly optimized by remarkable model capacity and better semantic relation preservation.

In summary, we propose a Cross-Modal Online Similarity function learning (CMOS) method to learn the asymmetric similarity function between heterogeneous data objects by preserving the relative semantic relation. The relative similarity is modeled as a set of bi-directional hinge loss constraints on the cross-modal training triplets. By measuring the semantic similarity in the label space, the relative similarity can also be applied to multi-label data. The overall online similarity function learning problem is formulated by the margin-based online Passive-Aggressive algorithm, and good scalability is gained in processing large scale datasets. We further extend our similarity learning model to combine multiple kernel functions to learn the similarity function in reproducing kernel Hilbert spaces. Experiments conducted on real-world datasets well demonstrate the effectiveness and efficiency of our approaches.

2. Related Work

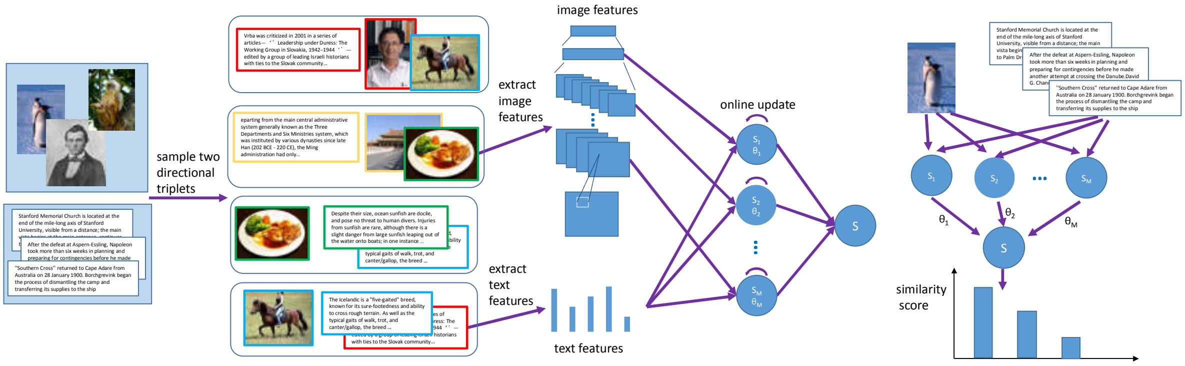
2.1. Cross Modal Correlation Learning

Many methods have been proposed for cross-modal correlation learning. Canonical correlation analysis (CCA) [25] is a classic method which learns the subspace that maximizes the correlation between two sets of aligned data items. Kernel canonical correlation analysis (KCCA)[11] is an extension of CCA using kernel trick. Several methods extend CCA to include supervised information, *e.g.*, GMA [29] and ml-CCA [24]. In addition, LCFS[32] and LGCFL[19] are based on linear regression to learn linear projections from feature spaces to label-based common space.

Information retrieval technique is employed by other methods, such as learning-to-rank [2, 10, 35, 41]. The bilinear similarity functions are learned by SSI [2], PAMIR [10] and RCCA [41] to minimize the pairwise ranking loss. In addition, a bi-directional list-wise ranking loss is optimized in Bi-CMSRM [35]. Another strand of research is on deep embedding method [39, 20, 33, 12] based on recent success in deep learning. Yan *et al.* [39] propose to learn the joint embeddings by deep canonical correlation analysis (DCCA)[1] which extends CCA with stacked nonlinear transformations. Wang *et al.* [33] propose to learn the joint embeddings by max-margin principle that combines cross-view ranking constraints with within-view neighborhood structure preservation constraints. Karpathy *et al.* [20] represent image regions using convolutional neural networks, words using bidirectional recurrent neural networks, and aligns fragments in an image and fragments in a text with a structured objective function.

2.2. Online Similarity Learning

In literature, a variety of algorithms have been proposed for online metric/similarity learning [28, 16, 3, 36] in single modality. Typically, a Mahalanobis distance metric [28, 16] is learned by online metric learning. While a bilinear metric [3, 10] is considered by some online similarity learning methods with the form $d(x_i, x_j) = x_i^T A x_j$, which can be seen as learning a linear transformation and taking inner product when A is positive semi-definite. OASIS [3] is an online similarity learning method which learns a bilinear similarity function for image retrieval. Xia *et al.* [36] extend bilinear similarity function to a more flexible nonlinear proximity function with multiple kernels to improve visual



(a) training (b) retrieval

Figure 1. Overview of the proposed method. The left figure shows the training phase. The right figure shows the retrieval phase by taking image-to-text retrieval as an example.

similarity search in CBIR. The online Passive-Aggressive learning technique is applied to learn the kernel-based similarity function for each individual kernel, and the Hedging online learning technique to learn the optimal combination weights of multiple kernels. However, there has been less study on cross-modal online similarity learning. PAMIR [10] is one that aims to rank images from text queries and can be kernelized. Since the model is based on mappings from the visual space to the text space, it only considers the single directional similarity constraints for model learning.

3. Approach

3.1. Notations

Without loss of generality, we use image and text modalities for illustration in this paper. Suppose that $\mathcal{V} = \{v_i\}_{i=1}^{N_v}$ is a set of images and $\mathcal{T} = \{t_i\}_{i=1}^{N_t}$ is a set of text documents. We overload notation by using v_i to denote both the image and its representation as a column vector $v_i \in \mathbb{R}^{d_v}$. Similarly, we use t_i to denote both the text and its representation as a column vector $t_i \in \mathbb{R}^{d_t}$. Let $r(v_i, t_j)$ denote the true pairwise semantic relevance between image v_i and text t_j. Relative similarity relationships are represented by two kinds of triplets $(v_i, t_i^+, t_i^-) \in \Pi^v$ and $(t_i, v_i^+, v_i^-) \in \Pi^t$. A triplet (v_i, t_i^+, t_i^-) indicates that $r(v_i, t_i^+) > r(v_i, t_i^-)$, and we call t_i^+ a positive example and t_i^- a negative example in this triplet for query v_i. Similar notations are applied on triplet (t_i, v_i^+, v_i^-). We denote a triplet by π_i which can be either kind of triplets.

3.2. Learning Bi-direction Relative Similarity

As we mentioned above, instead of explicitly learning the common space, we propose to learn a similarity function $s(v_i, t_j)$ to produce similarity score between image and text. We model the cross-modal similarity by relative similarity. The reason is two-folds. First, we need to sort the similarity score between database entries and the queries in retrieval context. Second, relative similarity may be constructed on label information, or in an unsupervised manner based on implicit users feedback.

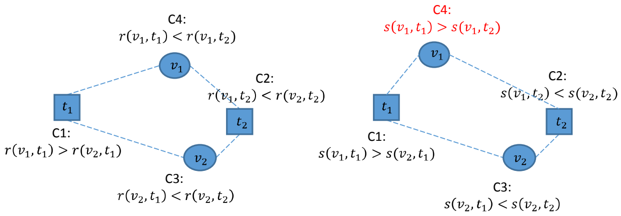

C4:
$r(v_1,t_1) < r(v_1,t_2)$
v_1
C2:
$r(v_1,t_2) < r(v_2,t_2)$
t_1
t_2
C1:
$r(v_1,t_1) > r(v_2,t_1)$
v_2
C3:
$r(v_2,t_1) < r(v_2,t_2)$
C4:
$s(v_1,t_1) > s(v_1,t_2)$
v_1
C2:
$s(v_1,t_2) < s(v_2,t_2)$
t_1
t_2
C1:
$s(v_1,t_1) > s(v_2,t_1)$
v_2
C3:
$s(v_2,t_1) < s(v_2,t_2)$

Figure 2. Drawback of singe direction similarity. Circles and squares represent images and texts, respectively. The left figure shows the true relationship of the four data objects, and the right figure shows the learned relationship of the four data objects. If the model is only learned to preserve relative similarity C1 and C2 in text-to-image direction, there is no guaranty that relative image-to-text similarity C3 and C4 can be satisfied. For example, v_1 to t_1 and t_2 similarity in C4 is not satisfied here.

There are two kinds of relative similarity for the similarity function, *i.e.*, the relative similarities of texts to an image and the relative similarities of images to a text. As show in Figure 2, bi-directional relative similarity constraints are indispensable for modeling the cross-modal semantic relation. Therefore, we expect the similarity function to satisfy the following two conditions simultaneously:

$$\begin{gathered} s(v_i, t_i^+) > s(v_i, t_i^-), \\ \forall v_i \in \mathcal{V}, t_i^+, t_i^- \in \mathcal{T} \text{ such that } r(v_i, t_i^+) > r(v_i, t_i^-), \\ s(v_i^+, t_i) > s(v_i^-, t_i), \\ \forall t_i \in \mathcal{T}, v_i^+, v_i^- \in \mathcal{V} \text{ such that } r(v_i^+, t_i) > r(v_i^-, t_i). \end{gathered}$$

We formulate the similarity function $s(v_i, t_j)$ as an asymmetric bilinear function:

$$s(v_i, t_j) = {v_i}^T \mathbf{W} t_j, \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d_v \times d_t}$ is not square, because the similarity function is defined between heterogeneous data space. The proposed definition is simple and flexible. Consequently, the similarity learning can fully capture the correlation patterns in cross-modal data. Note that this bilinear function can be regarded as a linear function on the joint feature of (v_i, t_j^T), so the multiplicative interactions between the two modalities can be measured [2]. The advantage over distance-based metric is that we do not have to specify the dimension of the learned latent common space.

To further improve the generalization performance of the similarity function learning, we introduce margins to the relative similarities as:

$$\begin{aligned} s(v_i, t_i^+) &> s(v_i, t_i^-) + 1, \\ s(v_j^+, t_j) &> s(v_j^-, t_j) + 1. \end{aligned} \tag{2}$$

To relax the training constraints for the nonfeasible case, we define hinge losses for the two directions as:

$$\begin{aligned} l_v(\mathbf{W}; v_i, t_i^+, t_i^-) &= \max\{0, s(v_i, t_i^-) - s(v_i, t_i^+) + 1\}, \\ l_t(\mathbf{W}; t_j, v_j^+, v_j^-) &= \max\{0, s(v_j^-, t_j) - s(v_j^+, t_j) + 1\}. \end{aligned} \tag{3}$$

On the whole, our goal is to minimize the empirical ranking loss with respect to the training data:

$$L(\mathbf{W}; D_{train}) = \sum_{\pi_i \in \Pi^v} l_v(\mathbf{W}; v_i, t_i^+, t_i^-) + \sum_{\pi_i \in \Pi^t} l_t(\mathbf{W}; t_i, v_i^+, v_i^-).$$

3.3. Online Learning Algorithm

To learn the similarity function efficiently, we use the Passive-Aggressive(PA) algorithm [5, 10, 3] to learn the similarity function. PA algorithm is a family of margin-based online learning algorithm closely related to stochastic gradient method. The update steps of PA is based on analytical solutions to simple constrained optimization problems.

By PA algorithm, we incrementally learn the weight matrix $\mathbf{W}$ with an iterative procedure. First, we initialize $\mathbf{W}$ as a zero matrix. Then, at each iteration i, we sample a triplet $\pi_i \in \Pi^v$ or $\pi_i \in \Pi^t$. If $\pi_i \in \Pi^v$, $\mathbf{W}_i$ is updated according to Eq.(4). Otherwise, we select $\mathbf{W}_i$ according to Eq.(5).

$$\mathbf{W}_i = \arg\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{W} - \mathbf{W}_{i-1}\|_F^2 + Cl_v(\mathbf{W}; v_i, t_i^+, t_i^-), \tag{4}$$

$$\mathbf{W}_i = \arg\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{W} - \mathbf{W}_{i-1}\|_F^2 + Cl_t(\mathbf{W}; t_i, v_i^+, v_i^-), \tag{5}$$

where $\|\cdot\|_F$ denotes Frobenius norm. In each iteration i, $\mathbf{W}_i$ is learned to reach a trade-off between minimizing the loss on the current triplet and remaining close to previous parameter $\mathbf{W}_{i-1}$. C is the aggressiveness parameter that controls the trade-off. In the iterative procedure, as we sample triplets from two directions, we update $\mathbf{W}$ from two directions, *i.e.*, the relative similarity of text-to-image and image-to-text.

By rewriting Eq.(4) as a constraint problem and defining the Lagrangian of the optimization problem as in [3], it can be shown that the solution of Eq.(4) is

$$\begin{aligned} \mathbf{W}_i &= \mathbf{W}_{i-1} + \tau_i \mathbf{V}_i, \\ \text{where } \mathbf{V}_i &= v_i \times (t_i^+ - t_i^-)^T, \\ \text{and } \tau_i &= \min\{C, \frac{l_v(\mathbf{W}_{i-1}; v_i, t_i^+, t_i^-)}{\|\mathbf{V}_i\|^2}\}. \end{aligned} \tag{6}$$

Similarly, the solution of Eq.(5) is:

$$\begin{aligned} \mathbf{W}_i &= \mathbf{W}_{i-1} + \tau_i \mathbf{V}_i, \\ \text{where } \mathbf{V}_i &= (v_i^+ - v_i^-) \times t_i^T, \\ \text{and } \tau_i &= \min\{C, \frac{l_t(\mathbf{W}_{i-1}; t_i, v_i^+, v_i^-)}{\|\mathbf{V}_i\|^2}\}. \end{aligned} \tag{7}$$

Algorithm 1 summarizes the proposed algorithm for Cross-Modal Online Similarity function learning (CMOS).

Algorithm 1 Cross-Modal Online Similarity Function Learning

```
Input: $\mathbf{W}_0 = \mathbf{0}$
Output: $\mathbf{W}$
  for $i = 1, \ldots, N$ do
    Sample triplet $\pi_i = (v_i, t_i^+, t_i^-)$ or $\pi_i = (t_i, v_i^+, v_i^-)$
    if $\pi_i \in \Pi^v$ then
      Update $\mathbf{W}_i$ as in Eq.(6)
    else
      Update $\mathbf{W}_i$ as in Eq.(7)
    end if
  end for
```

3.4. Online Multiple Kernel Learning

In real-world applications, the feature dimensions are usually prohibitively high. For example, features extracted from the middle layers of CNN are with more than 5K dimensions. We kernelize [5, 36] our model to deal with the curse-of-dimensionality problem in cross-modal similarity learning. It is easy to prove that $\mathbf{W}$ can be expressed as a linear combination of tensor products between training images and training texts. This proof is performed by induction over the iterations of our training procedure. Initially, $\mathbf{W} = \mathbf{0}$. so $\mathbf{W}$ can be seen as a linear combination of tensor products between training images and training texts with coefficient 0. Now we assume that the property is preserved at iteration $i - 1$. As we can see from the above subsection, $\mathbf{W}$ keeps the same when the selected triplet satisfies the constraint, or $\mathbf{W}$ is updated by adding times of tensor product between samples when the selected triplet violates the constraint, *i.e.*, $\mathbf{W}_i = \mathbf{W}_{i-1} + \tau_i \mathbf{V}_i$, where $\mathbf{V}_i$ is a tensor product between samples. Thus, at iteration i, the property is still preserved on $\mathbf{W}_i$. Therefore, $\mathbf{W}$ can be expressed as follows:

$$\mathbf{W} = \sum_{\pi_i \in \Pi^v} \tau_i v_i (t_i^+ - t_i^-)^T + \sum_{\pi_i \in \Pi^t} \tau_i (v_i^+ - v_i^-) t_i^T. \tag{8}$$

Given two new samples v and t, the similarity function can be calculated as:

$$s(v,t) = \sum_{\pi_i \in \Pi^v} \tau_i v^T v_i (t_i^+ - t_i^-)^T t + \sum_{\pi_i \in \Pi^t} \tau_i v^T (v_i^+ - v_i^-) t_i^T t.$$

Replacing inner product with kernel functions, the similarity function can be rewritten as:

$$s(v,t) = \sum_{\pi_i \in \Pi^v} \tau_i k^v(v, v_j)(k^t(t_j^+, t) - k^t(t_j^-, t)) + \sum_{\pi_i \in \Pi^t} \tau_i (k^v(v, v_j^+) - k^v(v, v_j^-)) k^t(t_j, t),$$

where $k^v(\cdot,\cdot)$ denotes kernel function for images, $k^t(\cdot,\cdot)$ denotes kernel function for texts. From the above equation, we can see that every kernelized similarity function needs a pair of kernel functions for image and text respectively. Any suitable kernel functions, *e.g.*, the Gaussian kernel function or polynomial kernel function, can be introduced. When learning the kernelized similarity function, we record coefficients τ_i and triplets sampled instead of recording $\mathbf{W}_i$. When calculating τ_i, we still use Eq.(6) and Eq.(7). But in these equations, we calculate the loss and $\|\mathbf{V}_i\|^2$ with kernel functions. With the kernelized similarity function, the loss incurred by samples can be easily calculated. $\|\mathbf{V}_i\|^2 = k^v(v_i, v_i)[k^t(t_i^+, t_i^+) - 2k^t(t_i^-, t_i^+) + k^t(t_i^-, t_i^-)]$ for $\pi_i \in \Pi^v$ and $\|\mathbf{V}_i\|^2 = k^t(t_i, t_i)[k^v(v_i^+, v_i^+) - 2k^v(v_i^-, v_i^+) + k^v(v_i^-, v_i^-)]$ for $\pi_i \in \Pi^t$.

The kernel function $k^v(\cdot,\cdot)$ is associated with an RKHS $\mathcal{H}_v$ endowed with an inner product $< \cdot,\cdot >_v$, and similarly $k^t(\cdot,\cdot)$ with $\mathcal{H}_t$ and $< \cdot,\cdot >_t$. Similar to [36], the kernelized similarity function can also be represented as $s_L(v,t) =< k^v(v,\cdot), L[k^t(t,\cdot)] >_{\mathcal{H}_v}$, where L is a linear operator mapping from Hilbert space $\mathcal{H}_t$ to $\mathcal{H}_v$.

With the development in representation learning, there are various features for image and text. Instead of selecting features by hand, we extend the above model to multiple kernel settings [36, 18]. Let $K = \{(k_j^v, k_j^t), j = 1, ..., M\}$ be a collection of M pairs of kernel functions. We want to learn the coefficients of linear combinations of the M pairs of kernels, while at the same time learn each similarity function. Let $f(v,t) = \sum_{j=1}^{M} \theta_j s_j(v,t)$, we consider the following optimization problem,

$$\min_{\theta \in \Delta} \min_{\{s_j\}_{j=1}^M} \frac{1}{2} \sum_{i=1}^{M} \|L_i\|_{HS}^2 + C(\sum_{\pi_i \in \Pi^v} l_v(f;\pi_i) + \sum_{\pi_i \in \Pi^t} l_t(f;\pi_i)),$$

where $\Delta = \{\theta \in \mathbb{R}_+^M | \theta^T e_M = 1\}$ is a simplex, e_M is a vector of ones, and $\|\cdot\|_{HS}$ is the Hilbert Schmidt norm of linear operator. Inspired by the success of the Hedging algorithm in learning combination weights [36, 18, 23], we apply the Hedging algorithm to learn the combination weights of multiple kernels in an online fashion. At every iteration, for each of the M pairs of kernels, *e.g.*, (k_j^v, k_j^t), we apply the above method to find the optimal coefficient for the kernelized similarity function with respect to kernel pair (k_j^v, k_j^t), and then apply the Hedging algorithm to update the combination weight by $\theta_j(i) = \theta_j(i-1)\beta^{z_j(i)}$, where $\beta \in (0,1)$ is a discounting parameter, and $z_j(i)$ equals to 1 when $s(v_i, t_i^+) - s(v_i, t_i^-) \le 0$ or $s(v_i^+, t_i) - s(v_i^-, t_i) \le 0$, and 0 otherwise. Algorithm 2 summarizes the proposed algorithm for Cross-Modal Online Multiple Kernel Similarity function learning(CMOMKS).

Algorithm 2 Cross-Modal Online Multiple Kernel Similarity Function Learning

```
Input: Kernel pairs $K = \{(k_j^v, k_j^t), j = 1, ..., M\}$, Discounting parameter $\beta$, Combination weights $\theta_j(0) = 1$.
Output: $f(v,t)$
  for $i = 1, \dots, N$ do
    Sample triplet $\pi_i = (v_i, t_i^+, t_i^-)$ or $\pi_i = (t_i, v_i^+, v_i^-)$
    for $j = 1, \dots, M$ do
      if $\pi_i \in \Pi^v$ then
        Compute $\tau_{ji}$ as in Eq.(6)
        Update $s_j(v,t)$ by adding term $\tau_{ji} k_j^v(v, v_i)$ $(k_j^t(t_i^+, t) - k_j^t(t_i^-, t))$
        if $s_j(v_i, t_i^+) - s_j(v_i, t_i^-) \le 0$ then
          $\theta_j(i) = \theta_j(i-1)\beta$
        end if
      else
        Compute $\tau_{ji}$ as in Eq.(7)
        Update $s_j(v,t)$ by adding term $\tau_{ji}(k_j^v(v, v_i^+) - k_j^v(v, v_i^-))$ $k_j^t(t_i, t)$
        if $s_j(v_i^+, t_i) - s_j(v_i^-, t_i) \le 0$ then
          $\theta_j(i) = \theta_j(i-1)\beta$
        end if
      end if
    end for
  end for
```

3.5. Mistake Bounds

We give mistake bounds for the above two algorithms in this subsection. We denote by $l_{vi}(l_{ti})$ the instantaneous loss suffered by our algorithm on iteration i. In addition, we denote by $l_{vi}^*(l_{ti}^*)$ the loss $l_{vi}(l_{ti})$ suffered by the arbitrary fixed predictor to which we are comparing our performance. Theorem 1 gives the mistake bound of Algorithm 1.

Theorem 1. *Let $\pi_i, \dots, \pi_N$ be a sequence of examples where $\pi_i \in \Pi^v$ or $\pi_i \in \Pi^t$. Assume $\|v_i(t_i^+ - t_i^-)^T\|_F^2$, when $\pi_i \in \Pi^v$, $\|(v_i^+ - v_i^-)t_i^T\|_F^2 \le R$, when $\pi_i \in \Pi^t$ for all i. Then, for any matrix $\mathbf{U} \in \mathbb{R}^{d^v \times d^t}$, the number of prediction mistakes made by Algorithm 1 on this sequence of examples is bounded from above by*

$$\max\{R, 1/C\}(\|\mathbf{U}\|_F^2 + 2C(\sum_{\pi_i \in \Pi^v} l_{vi}^* + \sum_{\pi_i \in \Pi^t} l_{ti}^*).$$

The proof can be easily obtained by extending Theorem 1 in [3]. Our learning algorithm is similar to [3], except that CMOS samples two kinds of triplets. Chechik *et al.* [3] have proved Theorem 1 by rewriting its formalisation as a linear classification problem. By folding the two kinds of triplets in this paper, we obtain the input vector with the same size for the linear classification problem.

Since we just replace inner product with kernel function for kernelized CMOS, the mistake bound is similar to CMOS but with the constraints becoming $k^v(v_i, v_i)(k^t(t_i^+, t_i^+) - 2k^t(t_i^+, t_i^-) + k^t(t_i^-, t_i^-)) \leq R$ and $k^t(t_i, t_i)(k^v(v_i^+, v_i^+) - 2k^v(v_i^+, v_i^-) + k^v(v_i^-, v_i^-)) \leq R$. Also we replace $\|\mathbf{U}\|_F^2$ with $\|T\|_{HS}^2$, where T is a linear operator mapping from the Hilbert space $\mathcal{H}_t$ to $\mathcal{H}_v$, and $\|\cdot\|_{HS}$ is the Hilbert Schmidt norm of linear operator.

For the multiple kernel settings, the number of mistakes T made by running the multiple kernel learning algorithm is denoted by $T = \sum_{i=1}^{N} I(\sum_{j=1}^{M} q_j(i - 1)z_j(i) \geq 0.5)$, where $I(x)$ is an indicator function. $\theta_i = \sum_j^M \theta_j(i), q_j(i) = \frac{\theta_j(i)}{\theta_i}$ defines the mixture of coefficient, and $z_j(i)$ indicates if triplet π_i is misclassified by the jth kernel similarity function. Following [8, 18, 36], we can get the upper bound of the number of prediction mistakes T. We define the optimal margin error as $g(k_j^v, k_j^t)$ for the kernel pair (k_j^v, k_j^t) with respect to a collection of training examples $\{\pi_i, i = 1, ..., N\}$, which satisfies $k_j^v(v_i, v_i)(k_j^t(t_i^+, t_i^+) - 2k_j^t(t_i^+, t_i^+) + k_j^t(t_i^-, t_i^+)) \leq R_j$, when $\pi_i \in \Pi^v$, and $k_j^t(t_i, t_i)(k_j^v(v_i^+, v_i^+) - 2k_j^v(v_i^+, v_i^+) + k_j^v(v_i^-, v_i^+)) \leq R_j$, when $\pi_i \in \Pi^t$. Theorem 2 gives the mistake bound of Algorithm 2.

Theorem 2. *Let $\pi_i, \ldots, \pi_N$ be a sequence of examples where $\pi_i \in \Pi^v$ or $\pi_i \in \Pi^t$. Assume $k_j^v(v_i, v_i)(k_j^t(t_i^+, t_i^+) - 2k_j^t(t_i^+, t_i^+) + k_j^t(t_i^-, t_i^+)) \leq R_j$, when $\pi_i \in \Pi^v$, $k_j^t(t_i, t_i)(k_j^v(v_i^+, v_i^+) - 2k_j^v(v_i^+, v_i^+) + k_j^v(v_i^-, v_i^+)) \leq R_j$, when $\pi_i \in \Pi^t$, for all i and j. Then, the number of prediction mistakes made by Algorithm 2 on this sequence of examples is upper-bounded by*

$$T \leq \frac{2\ln(1/\beta)}{1-\beta} \min_{1 \leq j \leq M} g(k_j^v, k_j^t) + \frac{2\ln M}{1-\beta}.$$

4. Experiments

4.1. Datasets

The Wiki dataset [25] contains 2,866 text-image pairs which are selected from the Wikipedia's featured articles collection. Each pair is labelled with one of ten semantic classes. 2173 pairs are taken as training set and 693 pairs are taken as testing set as in [25]. We use the publicly available 10-dim LDA text features [25], and extract CNN features using Caffe [17] with the pre-trained architecture learned on ImageNet. Features from 'conv2', 'conv5', 'fc6', 'fc7', 'prob' layers are extracted. After concatenating these CNN features and performing PCA by preserving 95% energy, we get 1648-dim image features for linear similarity learning and other linear similarity learning baseline approaches.

The Pascal VOC 2007 dataset [7] consists of annotated consumer photographs collected from Flickr. 399-dim tag occurrence features provided by [14] are used for text representations. 3394-dim CNN image features are extracted from concatenated CNN representations after PCA. For label representations, we use the groundtruth 20-class annotations of the images. The original train-test split provided in the dataset is used for training and testing. After removing images without tags, we get a training set with 5000 images and a test set with 4919 images.

The NUS-WIDE dataset [4] consists of images from Flickr accompanied with rich tags. In our experiment, We sample 10000 and 5000 image-text pairs which have 10 class labels with the largest number of images as training set and testing set, respectively. Texts are represented by 1000-dim tag occurrence vectors. We perform PCA on the concatenated CNN features by preserving 90% energy and get 3561-dim image features for linear similarity learning. 10-dim category indicator vectors are treated as ground-truth class labels.

4.2. Experimental settings

We compare our methods with CCA [25], PLS [26], GMLDA [29], Bi-CMSRM [35], SSI [2], LCFS [32], ml-CCA [24], KCCA [11] and KGMLDA [29]. For single-class methods on multi-label datasets (GMLDA and LCFS on the Pascal and NUS-WIDE dataset), the category of every training pair is provided with a randomly selected category from its multiple labels. We use the reduced concatenated CNN feature by PCA as the visual feature for each image.

In our methods, triplets sampling is needed for every iteration. In the training procedure, triplets in two directions are used alternatively. On the Wiki dataset, the positive example is the one in the same class with query, and the negative example is the one in different classes with query. On the Pascal and NUS-WIDE dataset, positive example shares more labels with query than negative example.

For kernel methods, Gaussian kernel function $k(x, y) = e^{-\frac{\|x-y\|^2}{\sigma^2}}$ are used. We conduct validation and select σ from $\{0.1, 1, 10, 100\}$. And we average kernel matrices from 'conv2', 'conv5', 'fc6', 'fc7', 'prob' layers for kernels in KCCA and KGMLDA. For our methods, the aggressiveness parameter C is set to 0.05 in all experiments, and the discount parameter β is set to 0.998. For the compared methods, we use the parameters' optimal settings tuned by a parameter validation process except for specification.

To evaluate the semantic consistency of our cross-modal similarity function, we perform bi-directional retrieval tasks in the experiments, *i.e.*, image-to-text retrieval and text-to-image retrieval. Images and texts are regarded to be relevant if they belong to the same class on the Wiki datast. While on the multi-label datasets Pascal and NUS-WIDE, images and texts are regarded to be relevant if they share at least one class label. Mean Average Precision(MAP) is used as the evaluation metric. Precision-recall curve is also presented to show the performance.

4.3. Results

The first three columns of Table 1 show the performance of cross-modal retrieval in terms of MAP on the Wiki dataset. From the experimental results, we can see that for primal methods, CMOS performs better than most of primal methods except SSI on image-to-text task. SSI is similar to CMOS. It learns a bilinear function by preserving relative similarity for each relative direction. However it has to set a learning rate which is not so intuitive like aggressiveness parameter and has no theoretical mistake bounds. Kernel methods outperform primal methods, because of non-linear transformation introduced by kernel functions. CMOMKS performs the best. By applying the Hedging method to combine different layers of CNN features, CMOMKS can select useful features and reduce the influence of useless features. The corresponding precision-recall curves are plotted in Figure 3(a) and Figure 3(b).

The middle three columns of Table 1 illustrate the performance of cross-modal retrieval in terms of MAP on the Pasal VOC 2007 dataset. From the experimental results, we can see that CMOS performs better than its primal counterparts and CMOMKS performs better than all baselines. Different from Wiki dataset, Pascal is a multi-label dataset. We can see that relative similarity can model the multi-label information well. The corresponding precision-recall curves are plotted in Figure 3(c) and Figure 3(d).

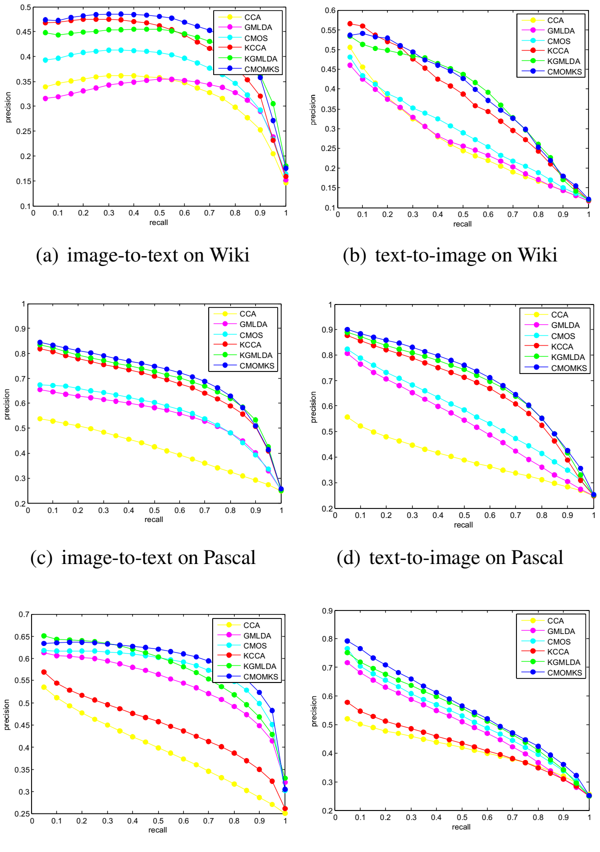
Results of cross-modal retrieval in terms of MAP on the NUS-WIDE dataset are showed in last three columns of Table 1. CMOS still outperforms the baseline methods and CMOMKS gets the best performance. Figure 3(e) and Figure 3(f) illustrate the corresponding precision-recall curves. We provide some hard examples in Figure 4, where our method made particular difference.

4.4. Further Analyses

To show how bi-directional learning works, we conduct experiments using triplets from different directions on the Pascal dataset. We sampled 10^6 training triplets used in all the three methods. Since there are image-text pairs in the Pascal dataset, it corresponds to 10^6 training triplets in Π_v and 10^6 training triplets in Π_t. For fair comparison, we trained single-directional methods for two times. Aggressiveness parameter C is fixed to 0.05 in all the three methods. Table 2 illustrates the results of the three methods. From the table, we can see that relative similarity in one direction can be useful for learning the relative similarity in other direction. Model trained with image-to-text triplets obtains good result in text-to-image retrieval task. Also it is easy to see that the information from two directions can be contained in one single model. In image-to-text retrieval, the performance of bi-directional model is better than image-to-text directional model. In text-to-image retrieval, the performance of bi-directional model is comparable with text-to-image directional model. On average, the bi-directional model performs the best, indicating that the triplets organized in two directions provide comprehensive information for similarity learning.

Method	img2txt	txt2img	average
image-to-text direction	0.579	0.586	0.583
text-to-image direction	0.577	0.603	0.59
bi-direction	0.586	0.600	0.593

Table 2. Performance of models trained with different directional triplets on the Pascal dataset.



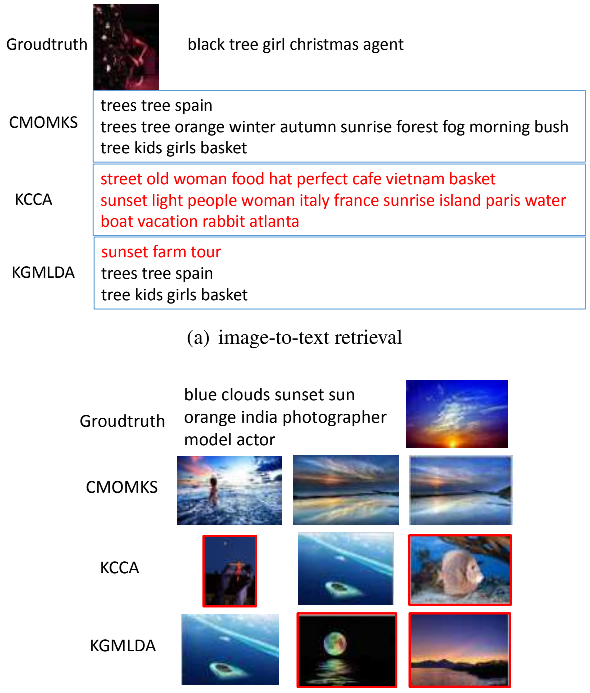
(a) image-to-text on Wiki
(b) text-to-image on Wiki
(c) image-to-text on Pascal
(d) text-to-image on Pascal
(e) image-to-text on NUS
(f) text-to-image on NUS

Figure 3. Performance of different methods on all benchmark datasets based on precision-recall curve.

We also examine the effect of the aggressiveness parameter C. We ran CMOS with different values of the parameter C on the Pascal dataset. Figure 5 depicts the retrieval

	Wiki			Pascal VOC 2007			NUS-WIDE		
Method	img2txt	txt2img	average	img2txt	txt2img	average	img2txt	txt2img	average
CCA	0.325	0.275	0.300	0.346	0.395	0.371	0.416	0.399	0.408
PLS	0.329	0.269	0.299	0.532	0.490	0.511	0.550	0.516	0.533
GMLDA	0.326	0.274	0.300	0.550	0.539	0.545	0.544	0.504	0.524
ml-CCA	0.343	0.293	0.318	0.584	0.572	0.578	0.546	0.501	0.524
LCFS	0.333	0.283	0.308	0.551	0.528	0.540	0.574	0.512	0.543
Bi-CMSRM	0.334	0.245	0.290	0.541	0.516	0.529	0.548	0.502	0.525
SSI	0.369	0.288	0.328	0.576	0.530	0.553	0.570	0.511	0.540
CMOS	0.368	0.298	0.333	0.586	0.600	0.593	0.578	0.528	0.553
KCCA	0.399	0.362	0.381	0.647	0.655	0.651	0.452	0.433	0.443
KGMLDA	0.414	0.359	0.386	0.675	0.676	0.676	0.578	0.544	0.561
CMOMKS	**0.434**	**0.388**	**0.411**	**0.709**	**0.707**	**0.708**	**0.597**	**0.565**	**0.581**

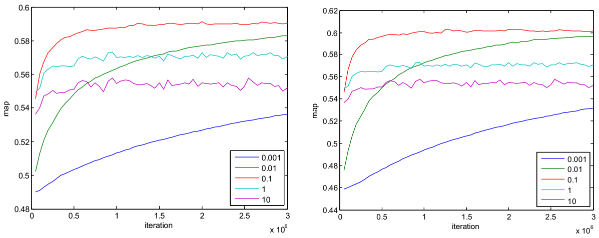
Table 1. Cross-modal retrieval performance of the proposed methods and the compared baselines on all benchmark datasets in terms of MAP.


Groudtruth
black tree girl christmas agent
CMOMKS
trees tree spain
trees tree orange winter autumn sunrise forest fog morning bush
tree kids girls basket
KCCA
street old woman food hat perfect cafe vietnam basket
sunset light people woman italy france sunrise island paris water
boat vacation rabbit atlanta
KGMLDA
sunset farm tour
trees tree spain
tree kids girls basket
Groudtruth
blue clouds sunset sun
orange india photographer
model actor
CMOMKS
KCCA
KGMLDA

(a) image-to-text retrieval

(b) text-to-image retrieval

Figure 4. Examples of top 3 results on the NUS-WIDE dataset. First row is the query and the corresponded cross-modal document. Red-colored texts and boxes indicate wrong results.



(a) image-to-text retrieval

(b) text-to-image retrieval

Figure 5. MAP with respect to the aggressiveness parameters C as a function of iterations on the Pascal dataset

performance with respect to C as a function of iterations. As can be seen from the figures, the value of the parameter C significantly influences the performances of the algorithms. Our model does not perform well with large value. When $C = 10$ and $C = 1$, MAP is low and vibrates as the number of iterations increases. With parameter $C = 0.1$, CMOS converges quickly, and a stable and good model can be obtained after sufficient number of iterations. Setting C to be a small value (*i.e.* 0.001) leads to a slow progress rate, since each online update changes the online hypothesis by a small amount.

5. Conclusions

We have proposed CMOS and its multiple kernel extension CMOMKS to learn a similarity function between heterogeneous data modalities by preserving relative similarity constraints from two directions. The online model is learned by the Passive-Aggressive algorithm. Multiple kernelized similarity function is further combined in CMOMKS, and the model is learned by the Hedging algorithm. Experimental results on three public cross-modal datasets have demonstrated that the proposed methods outperform state-of-the-art approaches. In future work, we will investigate to enhance scalability by accelerating the kernel calculation in online similarity learning process, and extend CMOMKS to deal with the situation where both modalities are represented by multiple representations.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China: 61672497, 61332016, 61620106009, 61650202, U1636214, 61572488, in part by National Basic Research Program of China (973 Program): 2015CB351802 and in part by Key Research Program of Frontier Sciences of CAS: QYZDJ-SSW-SYS013.

References

[1] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *ICML (3)*, pages 1247–1255, 2013.

[2] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, O. Chapelle, and K. Weinberger. Learning to rank with (a lot of) word features. *Information retrieval*, 13(3):291–314, 2010.

[3] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *JMLR*, 11:1109–1135, 2010.

[4] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: a real-world web image database from national university of singapore. In *CIVR*, page 48. ACM, 2009.

[5] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. *JMLR*, 7:551–585, 2006.

[6] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, 2007.

[7] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[8] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[9] A. Globerson and S. T. Roweis. Metric learning by collapsing classes. In *NIPS*, pages 451–458, 2005.

[10] D. Grangier and S. Bengio. A discriminative kernel-based approach to rank images from text queries. *TPAMI*, 30(8):1371–1384, 2008.

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

[12] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan. Cross-modal retrieval via deep and bidirectional representation learning. *IEEE Transactions on Multimedia*, 18(7):1363–1377, 2016.

[13] Y. Hua, S. Wang, S. Liu, Q. Huang, and A. Cai. Tina: Cross-modal correlation learning by adaptive hierarchical semantic aggregation. In *ICDM*, pages 190–199. IEEE, 2014.

[14] S. J. Hwang and K. Grauman. Accounting for the relative importance of objects in image retrieval. In *BMVC*, volume 1, page 5, 2010.

[15] P. Jain, B. Kulis, J. V. Davis, and I. S. Dhillon. Metric and kernel learning using a linear transformation. *JMLR*, 13(Mar):519–547, 2012.

[16] P. Jain, B. Kulis, I. S. Dhillon, and K. Grauman. Online metric learning and fast similarity search. In *NIPS*, pages 761–768, 2009.

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[18] R. Jin, S. C. Hoi, and T. Yang. Online multiple kernel learning: Algorithms and mistake bounds. In *Algorithmic Learning Theory*, pages 390–404. Springer, 2010.

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

[20] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[21] Z. Kuang and K.-Y. K. Wong. Relatively-paired space analysis: Learning a latent common space from relatively-paired observations. *IJCV*, 113(3):176–192, 2015.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[23] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang. Hedged deep tracking. In *CVPR*, pages 4303–4311, 2016.

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

[25] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260. ACM, 2010.

[26] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *Subspace, latent structure and feature selection*, pages 34–51. Springer, 2006.

[27] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. *NIPS*, page 41, 2004.

[28] S. Shalev-Shwartz, Y. Singer, and A. Y. Ng. Online and batch learning of pseudo-metrics. In *ICML*, page 94. ACM, 2004.

[29] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, pages 2160–2167. IEEE, 2012.

[30] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.

[31] J. Wang, H. T. Do, A. Woznica, and A. Kalousis. Metric learning with multiple kernels. In *NIPS*, pages 1170–1178, 2011.

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

[33] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, pages 5005–5013, 2016.

[34] S. Wang, S. Jiang, Q. Huang, and Q. Tian. Multi-feature metric learning with knowledge transfer among semantics and social tagging. In *CVPR*, pages 2240–2247. IEEE, 2012.

[35] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang. Cross-media semantic representation via bi-directional learning to rank. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 877–886. ACM, 2013.

[36] H. Xia, S. C. Hoi, R. Jin, and P. Zhao. Online multiple kernel similarity learning for visual search. *TPAMI*, 36(3):536–549, 2014.

[37] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. *NIPS*, 15:505–512, 2003.

[38] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81, 2015.

[39] F. Yan and K. Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, pages 3441–3450, 2015.

[40] F. Yan, K. Mikolajczyk, and J. Kittler. Multiple kernel learning via distance metric learning for interactive image retrieval. In *International Workshop on Multiple Classifier Systems*, pages 147–156. Springer, 2011.

[41] T. Yao, T. Mei, and C.-W. Ngo. Learning query and image similarities with ranking canonical correlation analysis. In *ICCV*, pages 28–36, 2015.