Multi modal representation learning and cross-modal semantic matching

Wang, X.

4

KERNEL-BASED MIXTURE MAPPING FOR IMAGE AND TEXT ASSOCIATION

This chapter is based on the following publication:

Du, Y., Wang, X., Cui, Y., Wang, H., Su, C. (2019). Kernel-Based Mixture Mapping for Image and Text Association. IEEE Transactions on Multimedia, 22(2), 365-379.

CHAPTER SUMMARY

This chapter addresses RQ4.

RQ4: How and with what quality can we model the semantic correlations between two different modalities?

Modeling the relationship between multi modal media, including images, videos, and text, can reduce the gap between the modalities and promote cross-media retrieval, image annotation, etc. In this chapter, we propose a new approach called kernel-based mixture mapping (KMM) to model the semantic correlations between web images and text. With this approach, we first construct latent high-dimensional feature spaces based on kernel theory to address the non-linearity of both the data distributions in the input spaces and the cross-model correlation. Second, we present a probabilistic neighborhood model to describe the spatial locality of semantics by assuming that proximate examples in feature spaces generally have the same semantics and a conditional model to describe cross-modal conditional dependency. Finally, we build a probabilistic mixture model to jointly model the spatial locality of semantics and the conditional dependency between different modalities. By combining nonlinear transformation and probabilistic models, KMM can address the non-linearity of cross-modal correlation, the complexity of the semantic distributions at the global scale, and the continuity of semantic distributions at the local scale. We present a hybrid optimization algorithm to find the solution of KMM based on expectation-maximization and sub gradient ascent; this algorithm avoids estimating the parameters of KMM in high-dimensional feature spaces and is proved to converge to an (local) optimal solution. We demonstrate the performance of KMM using for public datasets.The experimental results show that our approach outperforms the compared methods when modeling the relationships between image and text.

With the rapid development of the Internet, there has been a massive explosion of multimedia content, such as text, image, audio and videos, on the web. These types of content usually coexist in a multimedia document and complement each other to express similar semantics. For example, an image provides a visual description of a concept, yet this description is usually incomplete. In contrast, text can accurately describe the abstraction of a concept, but it is not intuitive. Consequently, joint exploitation of the full information from different modalities could facilitate accurate content interpretation. Currently, many real-world internet applications, such as cross-media retrieval [90, 91, 92], image caption or summary generation [93], image annotation [94, 95, 96] and information recommendation [97], involve multimodal data. For these applications, the relationship between modalities needs to be considered. Many previous studies focused on the understanding of the unimodal scenario, in which data are homogeneously represented and similarity is measured in a single feature spaces. However, different data modalities are associated with different metric spaces, and thus similarity cannot be measured directly between heterogeneous modalities. The vastly different representations derived from heterogeneous modalities make it very challenging to associate signals across these modalities.

The work related to the semantic correlation mining of heterogeneous media can be categorized into the following four main classes: 1) linear/non-linear mapping [98], [99], such as canonical correlation analysis (CCA), 2) probabilistic models, such as probabilistic latent semantic analysis (PLSA) [100], 3) graph-based correlation propagation methods [92], [101], and 4) deep learning-based methods [102],[103]. In [43], the authors presented an approach called mixture of local linear mapping (MLLM) to cross-modal semantic correlation modeling. MLLM considers that close examples in a local region generally represent a uniform concept and are supposed to be mapped to another modality based on a linear model, and then combines multiple linear mapping models to represent the relationships between different modalities on the whole data distribution. However, MLLM cannot address the non-linearity of data distributions and cross-media correlations very well.

In this chapter, we first analyze the ineffectiveness of linear mapping models and then propose a novel approach, called kernel-based mixture mapping (KMM), to model the semantic association between text and images. Similar to our previous method MLLM, KMM considers that the data in a local region of the input spaces follow a local mapping model and uses a mixture of local mapping models to substitute a more complex nonlinear mapping. In KMM, we introduce a probabilistic neighborhood model to accurately describe how data in a local region follow the corresponding local mapping model. To address the nonlinearity of data distributions and cross-media correlations, KMM first transforms the textual and visual data from the input spaces into two latent high-dimensional spaces by nonlinear feature space mapping functions, and then constructs the mapping model between both modalities in the latent features spaces. The smoothness and sparseness of the parameters are introduced to enhance the generaliza-

tion of models and the fitness between models and data. We present a hybrid optimization algorithm based on expectation-maximization (EM) and subgradient ascent to find the solution of KMM; the parameters are estimated using kernel theory to avoid the explicit representation of both feature spaces.

In summary, our contributions are three-fold: 1) We analyze the ineffectiveness of linear models and reveal that linear models' prediction is close to a zero vector for cross-media retrieval due to the linear uncorrelation between images and text at the global scale in feature space. 2) We present a parameterized model-driven approach, called KMM, to model cross-modal association. KMM provides a kernel-based probabilistic mixture model to describe the distribution that cross-modal data need to follow and addresses the complexity of the semantic distribution at the global scale, its continuity at the local scale, and the non-linearity in the mapping of different modalities. 3) We introduce a hybrid optimization algorithm based on the frameworks of EM and subgradient ascent and prove its convergence to an (local) optimum. The optimization algorithm overcomes the difficulty in estimating the parameters of the KMM model because our model does not consist of explicit inner products for being replaced directly by kernel functions.

The rest of this chapter is organized as follows. Section 4.1 presents a brief overview of related work. Section 4.2 breifly analyzes the ineffectiveness of linear mapping models. Section 4.3 describes our KMM approach to the modeling of image and text association. Section 4.4 presents the optimization, algorithm and analysis for KMM. Section 4.5 provides the experimental results, and Section 4.6 concludes the chapter.

4.1. RELATED WORK

Studies related to cross-media modeling can be divided into four main classes.

(1) **Linear or nonlinear mapping**. This class of methods builds a linear or nonlinear (closed-form) transformation model between heterogeneous input spaces or from both input spaces to a latent semantic space where similarity is measured. Grangier and Bengio [104] proposed a linear discriminant approach for cross-modal retrieval by linearly transforming one modality to the other and extended the linear model to a nonlinear one through the kernel trick. Jiang and Tan [105] presented a vague linear transformation to measure the information similarity between visual and textual modalities through a set of predefined domain-specific information categories. There are some other approaches that transform both modalities into a common space, which can be constructed based on CCA [90, 106, 107], matrix factorization [108, 109], or by preserving a certain structure of data [110]. The similarity between multiple modalities can be measured in the common space. Tang et al. [111] presented a cross-space affinity model that was learned with an optimization problem, where the restriction of exact correspondences between different modalities was relaxed to their relative similarities. In addition, some researchers proposed mixture models to describe the relationship between two sources of data. In [112], Deleforge et al. introduced a model

called Gaussian locally-linear mapping (GLLiM) for high-dimensional regression. Different from [43], GLLiM model aims to solve the inverse regression problem. Hannah et al. [113] presented a more general regression model by introducing generalized linear models. To handle diverse content more appropriately, Hua et al. [114] presented a method called TINA that built a set of local linear projections for each modality and then measured the relations of pairs of local models for different modalities. To address nonlinearity of data distributions, Zhang et al. [115] and Xu et al. [116] introduced kernel mapping in data representation.

(2) **Probabilistic methods**. Probabilistic methods generally aim to maximize the probabilities that the data of one modality can be generated for the given inputs of the other modality. Jeon et al. [117] proposed an approach to annotating and retrieving images that directly modeled the joint distributions over blobs in images and words in text. Different from [117], Monay and Pere [100] computed the joint distributions over images and text based on PLSA by introducing a latent semantic variable. Feng and Lapata [118] proposed an approach to image captioning based on the latent Dirichlet allocation (LDA) model and generated the keywords of captions by maximizing the posterior probabilities given the image and its corresponding textual documents. Zhang et al. [119] supposed that features of images and text are independently generated by a certain concept and modeled cross-media relationships under the Bayesian framework. To improve learning performance, Wu et al. [120] incorporated unlabeled data in the training process of image retrieval and learned the model by maximizing the joint probabilities of labeled and unlabeled data with the discriminant-EM algorithm. To relax the restriction discussed in many studies regarding full correspondence between modalities, Jia et al. [121] proposed a method for analyzing the semantic correlation between modalities based on a Markov random field of topic models for realistic scenarios, where a narrative text is only loosely related to an image. Different from the above studies, Pham et al. [122] presented a method for learning fine-grained relationships between images and text, i.e., the correspondences between the keywords in text and the visual regions in images, based on EM algorithm.

(3) **Graph-based correlation propagation**. Generally, graph-based methods model multimedia with each document as a vertex and the relationship between documents as an edge, and propagate the correlation information to learn the cross-modality similarity over the graph [92]. Zhai et al. [101] constructed a kNN graph for each modality and performed cross-media retrieval by determining whether the examples from different modalities have the same label or not. Lin et al. [123] presented a PLSA-based aspect model to measure the inter-correlation between different modalities and intra-correlation in the same modality, and then constructed a multi-modal propagation network for cross-media retrieval. Lazaridis et al. [124] presented a novel framework based on kNN graphs for multimodal search of rich media objects, in which Laplacian eigenmaps were employed to merge low-level descriptors and create a new low-dimensional multimodal feature space. Xue et al. [125] proposed a graph-based approach that contained

two processes of semantic correlation computing for modeling the semantic correlation between web images and text. In the work, information propagation was jointly driven by the local semantics of visual blobs or words and the global semantics of documents. In [126], a multiple graph-based multi-label learning framework was proposed for image annotation problem, in which the visual content of images, semantic correlation of tags and the prior information provided by users were simultaneously considered. The multi-graph strategy was also used in [127], where the authors jointly modeled the intra-modal local topology structures of each graph constructed on one modality and the inter-modal local topology structures to obtain the final common embedding space for multiple modalities.

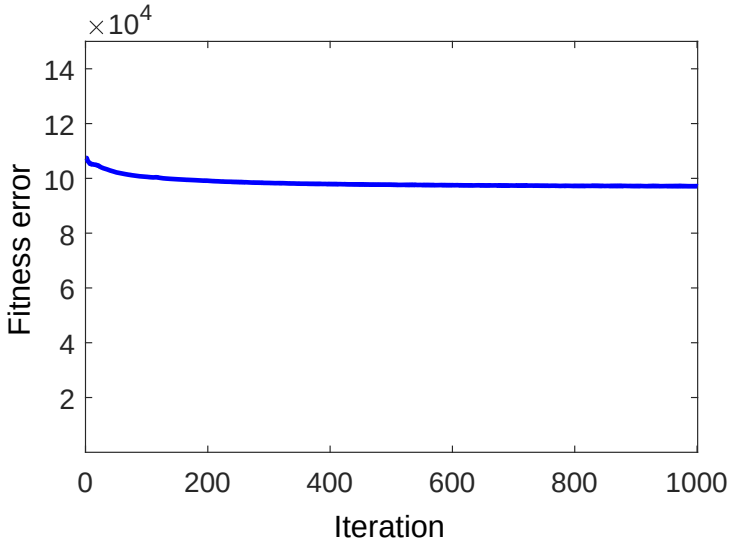
(4) **Deep learning-based methods**. In general, deep learning-based methods jointly map different modalities into an embedding space using deep networks and measure similarity in this space. Deep CCA [128] is a representative approach to cross-media correlation modeling, which represents each modality with a deep network and measures the similarity based on CCA. Different from deep CCA, Wang et al. [129, 103] measured the similarity between different modalities based on cross-view ranking constraints or the element-wise product. Eisenschtat and Wolf [102] presented a bidirectional neural network architecture for matching images and text, in which two tied neural network channels were used to project both views into a common, maximally correlated space using Euclidean loss. To make cross-modal correlation modeling more precise, Peng et al. [130] fused coarse-grained instances and fine-grained patches and learned the relationships between images and text based on the constraints of the intra-modality semantic category and the inter-modality pairwise similarity. To make an efficient retrieval, Hong et al. [131] presented a novel joint semantic-visual space by leveraging visual descriptors to narrow the semantic gap and provided an efficient online multimedia service. In addition to image-text association modeling, Wang et al. [132] focused on making correlations between movies and text and proposed a novel model called layered memory network, which can encode the temporal alignment between sentences and frames inside movie clips. Most of the deep learning-based methods model the relationships between different modalities by parameter tuning in the representation process. Different from these methods, we build an explicit probabilistic model to describe the cross-modality relation based on the representation from deep networks.

4.2. LINEAR MODELS AND THE INEFFECTIVENESS

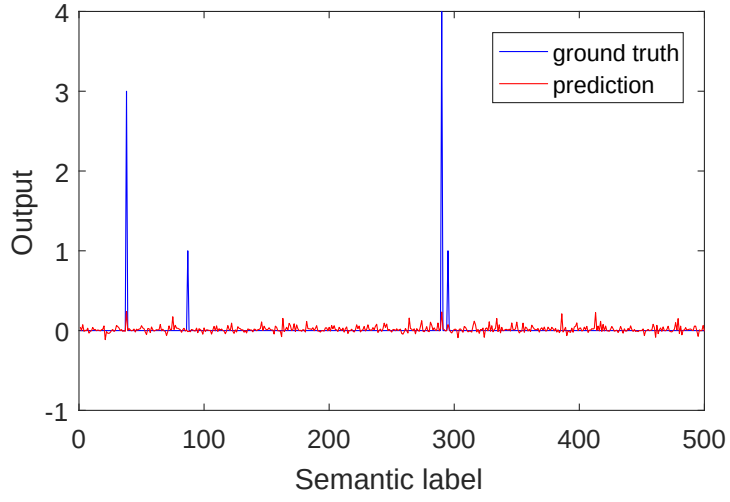
In general, there is a natural correspondence between visual space and textual space. Let

$$\mathscr{M} : \mathbf{R}^T \rightarrow \mathbf{R}^I \tag{4.1}$$

be an invertible map from the textual space to the visual space. Similarly, given a query of visual image $\mathbf{x}^I \in \mathbf{R}^I$, its corresponding textual sample in textual space can be achieved with the inverse of $\mathscr{M}$, i.e., $\mathscr{M}^{-1}(\mathbf{x}^I)$.



(a)



(b)

Figure 4.1: An example of linear transformation from textual spaces to visual spaces using the Corel5K dataset: (a) the decrease of fitness error with iteration and (b) the comparison between the prediction (red curve) and the ground-truth image (blue curve) for a certain textual input. In Fig. 4.1(b), images are represented by the bag of visual words (BoVW) model (500 visual words), and the y-axis denotes the value of each entry of feature vectors for the prediction or corresponding ground truth given a textual input.

Many previous models for mapping the heterogeneous modalities are constructed as linear models [90, 106, 105]. Jiang and Tan [105] transformed text (or images) to the other modality with a linear model and computed the similarity between the ground truth and the corresponding prediction:

$$\hat{\mathbf{x}}^I = \mathbf{M}_{CI}\mathbf{M}_{TC}\mathbf{x}^T \ , \tag{4.2}$$

where $\hat{\mathbf{x}}^I$ is the prediction in the visual space, and $\mathbf{M}_{TC}$ and $\mathbf{M}_{CI}$ are the transformation matrices from textual spaces to concept spaces and from concept spaces to visual spaces, respectively. The similarity can be measured using Euclidean distance in both spaces:

$$d_F(\mathbf{x}^I, \mathbf{x}^T) = \|\mathbf{x}^I - \mathbf{M}_{CI}\mathbf{M}_{TC}\mathbf{x}^T\|_2 \ . \tag{4.3}$$

4

Actually, images and text originate from two completely different systems. Furthermore, visual features and textual features are complicated and nonlinearly distributed in their respective spaces. Consequently, constructing a map between both spaces with a linear model is intuitively inaccurate. Fig. 4.1 illustrates an example of linear transformation from textual spaces to visual spaces using the Corel5K dataset. From Fig. 4.1(a), we find that the fitness error decreases by only approximately 10% through iterative optimization. Fig. 4.1(b) shows that the prediction result in visual space is similar to a random noise around 0 along the semantic label dimension and has a large difference from the ground truth. Theorem 1 provides a theoretical analysis.

Theorem 1. *If $\mathbf{x}^T$ and $\mathbf{x}^I$ are linearly uncorrelated, the solution to Eq.4.2 with the minimization of the distance shown in Eq.4.3 over all data is a zero vector that is independent of the distribution of $\mathbf{x}^I$.*

Proof. Let $\mathbf{x}^I = (\mathbf{x}_1^I, \mathbf{x}_2^I, \cdots, \mathbf{x}_N^I)$ and $\mathbf{x}^T = (\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_N^T)$ be the data matrices for images and text, respectively. Without loss of generality, we assume that $\mathbf{x}^I$ and $\mathbf{x}^T$ have a mean of zero. When $\mathbf{x}^T$ and $\mathbf{x}^I$ are linearly uncorrelated, the correlation coefficient can be computed as follows:

$$\begin{aligned} \rho &= \frac{tr(\mathbf{C}_{TI}\mathbf{C}_{TI}^T)}{\sqrt{tr(\mathbf{C}_{TT}\mathbf{C}_{TT}^T)tr(\mathbf{C}_{II}\mathbf{C}_{II}^T)}} \\ &= 0 \ , \end{aligned} \tag{4.4}$$

where $\mathbf{C}_{TI} = \mathbf{X}^I\mathbf{x}^{T\prime}$, $\mathbf{C}_{TT} = \mathbf{X}^T\mathbf{x}^{T\prime}$ and $\mathbf{C}_{II} = \mathbf{X}^I\mathbf{x}^{I^T}$. In this chapter, $\mathbf{x}^{T\prime}$ means transpose of the matrix $\mathbf{x}^T$. Thus, $tr(\mathbf{C}_{TI}\mathbf{C}_{TI}^T) = \|\mathbf{C}_{TI}\|_F^2 = 0$. By minimizing the distance in Eq.4.3, the following is attained:

$$\begin{aligned} \hat{\mathbf{x}}^I &= \mathbf{M}_{CI}\mathbf{M}_{TC}\mathbf{x}^T \\ &= \mathbf{X}^I\mathbf{x}^{T\prime}(\mathbf{x}^T\mathbf{x}^{T\prime})^{-1}\mathbf{x}^T \\ &= \mathbf{C}_{TI}\mathbf{C}_{TT}^{-1}\mathbf{x}^T \ . \end{aligned} \tag{4.5}$$

Then, $\|\hat{\mathbf{x}}^I\| \leq \|\mathbf{C}_{TI}\|_F\|\mathbf{C}_{TT}^{-1}\mathbf{x}^T\| = 0$, and thus $\hat{\mathbf{x}}^I = 0$. □

4.3. PROPOSED MODEL

4.3.1. LOCAL LINEAR MAPPING

In this approach, we write the map as $\mathscr{M} : X \rightarrow Y$. Without loss of generality, we let $X = \mathbf{R}^T$ and $Y = \mathbf{R}^I$. We consider that the map from a local region of X to Y can be described by a linear model due to the simplicity of the local data distribution. We characterize the linear mapping model $\mathscr{M}$ over the local region by the concatenation of two matrices as follows:

$$\begin{aligned} \mathbf{y}_i &= \hat{\mathbf{y}}_i + \varepsilon_i \\ &= \mathbf{W} \cdot \mathbf{V}\mathbf{x}_i + \varepsilon_i \ , \end{aligned} \tag{4.6}$$

where $\mathbf{x}_i \in X$ denotes an input (or a query), $\mathbf{y}_i, \hat{\mathbf{y}}_i \in Y$ are the corresponding ground-truth output and the prediction in the other modality, respectively, $\mathbf{V}$ is the transformation matrix from the input space to a latent semantic space, $\mathbf{W}$ is the transformation matrix from the semantic space to the output space, and ε_i denotes the fitness error. In our work, we assume the fitness error ε_i follows a normal distribution with zero mean and covariance matrix Σ. Given the model $\mathscr{M}$ and an input $\mathbf{x}_i$, the probability distribution of the ground-truth output $\mathbf{y}_i$ is formulated as follows:

$$\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathscr{M}) = \frac{1}{\sqrt{(2\pi)^{d_y}|\Sigma|}} e^{-\frac{1}{2}d(\mathbf{y}_i, \mathbf{x}_i)} \ , \tag{4.7}$$

where $d(\mathbf{y}_i, \mathbf{x}_i) = (\mathbf{y}_i - \mathbf{W}\mathbf{V}\mathbf{x}_i)^T \Sigma^{-1} (\mathbf{y}_i - \mathbf{W}\mathbf{V}\mathbf{x}_i)$.

As analyzed above, we consider that a set $\mathscr{R}_m$ of close examples in a local region indexed by m has uniform semantics and approximately follows one cross-media mapping model. Intuitively, the data near the centroid of $\mathscr{R}_m$ follow the mapping model with high confidence, and those far from the centroid follow with low confidence. We then characterize the confidence with a neighborhood model $K_{\mathbf{H}}(\mathbf{x} - \mu)$ with a symmetric positive definite $d_x \times d_x$ bandwidth matrix $\mathbf{H}$, where μ is the centroid of the local region and

$$K_{\mathbf{H}}(\mathbf{x} - \mu) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}(\mathbf{x} - \mu)) \ . \tag{4.8}$$

$K(\mathbf{x})$ is a bounded function with compact support satisfying [133]

$$\begin{aligned} &\int_{\mathbf{R}^{d_x}} K(\mathbf{x}) d\mathbf{x} = 1 \quad \lim_{||\mathbf{x}|| \rightarrow \infty} ||\mathbf{x}||^{d_x} K(\mathbf{x}) d\mathbf{x} = 0 \\ &\int_{\mathbf{R}^{d_x}} \mathbf{x} K(\mathbf{x}) d\mathbf{x} = 0 \quad \int_{\mathbf{R}^{d_x}} \mathbf{x}\mathbf{x}^T K(\mathbf{x}) d\mathbf{x} = c_K \mathbf{I} \ , \end{aligned} \tag{4.9}$$

where c_K is a constant. A Gaussian function with a zero-mean vector and an identity covariance matrix satisfies such constraints in Eq.4.9. We use $K_{\mathbf{H}_m}(\mathbf{x} - \mu_m)$ to describe the probability or confidence of the data that follow the mapping model $\mathscr{M}_m$ over $\mathscr{R}_m$, i.e., $\Pr(\mathbf{x}_i|\mathscr{M}_m)$.

The joint probability of the pair $(\mathbf{x}_i, \mathbf{y}_i)$ generated by the model $\mathscr{M}_m$ is:

$$\begin{aligned}\Pr(\mathbf{x}_i, \mathbf{y}_i|\mathscr{M}_m) &= \Pr(\mathbf{x}_i|\mathscr{M}_m)\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathscr{M}_m) \\ &= K_{\mathbf{H}_m}(\mathbf{x}_i - \mu_m)\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathscr{M}_m)\ ,\end{aligned} \tag{4.10}$$

where μ_m and $\mathbf{H}_m$ denote the centroid (replaced by the mean vector in computing) and bandwidth matrix, respectively, of the local region $\mathscr{R}_m$ that $\mathbf{x}_i$ belongs to.

An alternative factorization of the joint probability shown in Eq.4.10 can be performed as follows:

$$\Pr(\mathbf{x}_i, \mathbf{y}_i|\mathscr{M}'_m) = \Pr(\mathbf{y}_i|\mathscr{M}'_m)\Pr(\mathbf{x}_i|\mathbf{y}_i, \mathscr{M}'_m)\ ,$$

where $\mathscr{M}'_m$ denotes a mapping model from Y to X. This factorization is related to the inverse regression [112], where $\mathbf{y}_i$ is considered as a regressor. In this case, we need to define a neighborhood model $K_{\mathbf{H}}(\mathbf{y} - \mu)$ to describe the probability of $\mathbf{y}_i$ that follows the mapping model, i.e., $\Pr(\mathbf{y}_i|\mathscr{M}'_m)$. Compared with Eq.4.10, this factorization will result in a high computational complexity because we need to compute $\Pr(\mathbf{y}_i|\mathscr{M}'_m)$ for all $\mathbf{y}_i$ in a dataset given a certain query $\mathbf{x}_i$.

4.3.2. KERNEL-BASED MIXTURE MAPPING

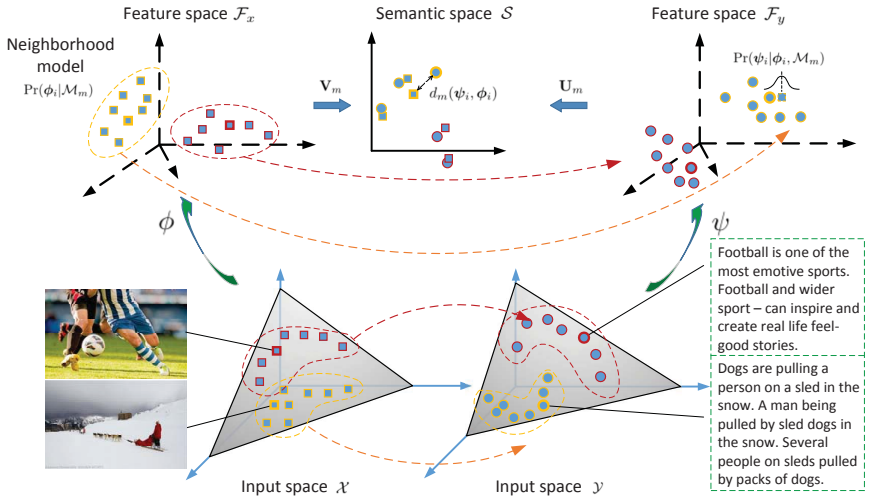
Due to the complexity of the data distribution, the map between two modalities may not follow the linear model in the original input space, and the local region in the input space cannot be depicted well by the expected Gaussian neighborhood model. Therefore, we formulate this problem in a high-dimensional latent feature space based on kernel theory. Let us consider $\phi : X \to \mathscr{F}_x$ and $\psi : Y \to \mathscr{F}_y$ that map the original input spaces into two feature spaces of dimensions d_ϕ and d_ψ, respectively, where both $\mathscr{F}_x$ and $\mathscr{F}_y$ are inner product spaces. Here, as shown in Fig. 4.2(a), we build a d_s-dimensional semantic space S by the linear transformation over both feature spaces. In the semantic space, it is easier to introduce the kernel theory and measure the similarity of two modalities. Similar to Eq.4.7, the map between two modalities can be represented by the following probabilistic model in the semantic space:

$$\Pr(\psi_i|\phi_i, \mathscr{M}_m) = \frac{1}{\sqrt{(2\pi)^{d_s}|\Sigma_m|}} e^{-\frac{1}{2}d_m(\psi_i, \phi_i)}\ , \tag{4.11}$$

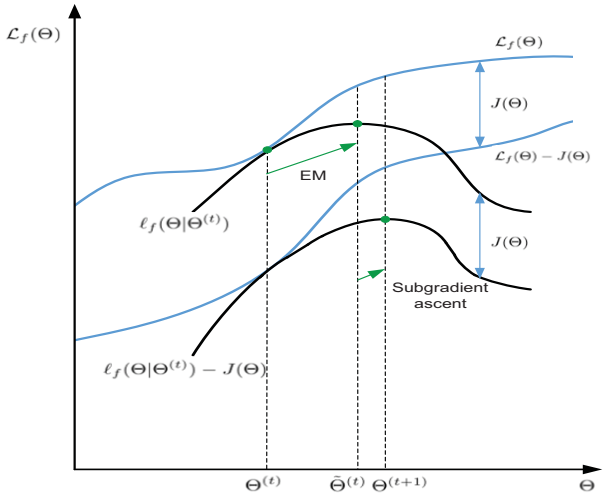
where $\phi_i \triangleq \phi(\mathbf{x}_i)$, $\psi_i \triangleq \psi(\mathbf{y}_i)$, and

$$d_m(\psi_i, \phi_i) = (\mathbf{U}_m\psi_i - \mathbf{V}_m\phi_i)^T\Sigma_m^{-1}(\mathbf{U}_m\psi_i - \mathbf{V}_m\phi_i). \tag{4.12}$$

where Σ_m denotes the covariance matrix of data points $\{\mathbf{U}_m\psi_i - \mathbf{V}_m\phi_i\}$ associated with the model $\mathscr{M}_m$ in semantic space and Σ_m^{-1} denotes the inverse. The distance $d_m(\cdot,\cdot)$ measured in S is achieved by the combination of features with matrices $\mathbf{U}_m$ and $\mathbf{V}_m$. Generally, the rows of $\mathbf{U}_m$ and $\mathbf{V}_m$ are located in the space spanned by the



(a)



(b)

Figure 4.2: Our approach. (a) The framework. The small squares and circles denote examples of images and text, respectively, located in input or feature spaces, and different colors indicate different local regions. In the input space, the local regions are supposed to follow a Gaussian neighborhood model in the feature space, while they do not follow this model in the input space. (b) Convergence analysis of hybrid optimization, which is introduced in Section 4.4 in detail.

columns of $\Psi = (\psi_i)$ and $\Phi = (\phi_i)$, respectively, i.e., $\mathbf{U}_m = \mathbf{A}_m \Psi^T$ and $\mathbf{V}_m = \mathbf{B}_m \Phi^T$. Eq.4.12 can be rewritten as:

$$d_m(\psi_i, \phi_i) = (\mathbf{A}_m K_{y,i} - \mathbf{B}_m K_{x,i})^T \Sigma_m^{-1} (\mathbf{A}_m K_{y,i} - \mathbf{B}_m K_{x,i}), \tag{4.13}$$

where each row of $\mathbf{A}_m$ and $\mathbf{B}_m$ denotes the coefficients with which the rows of $\mathbf{U}_m$ and $\mathbf{V}_m$ can be linearly reconstructed by the data points $\{\psi_i\}$ and $\{\phi_i\}$, respectively, and $K_{x,i}$ and $K_{y,i}$ denote the i-th column of kernel matrices $\mathbf{K}_x = (\phi_k \cdot \phi_l)$ and $\mathbf{K}_y = (\psi_k \cdot \psi_l)$, respectively. Based on the kernel theory [134], we can choose nonlinear kernel functions $f_\phi : X \times X \to \mathbf{R}$ and $f_\psi : Y \times Y \to \mathbf{R}$, which should follow Mercer's condition, to satisfy $f_\phi(\mathbf{x}_k, \mathbf{x}_l) = \phi_k \cdot \phi_l$ and $f_\psi(\mathbf{y}_k, \mathbf{y}_l) = \psi_k \cdot \psi_l$. Therefore, we can achieve the kernel matrix in input space instead of in feature space by choosing appropriate kernel functions.

4

The neighborhood model in the feature space $\mathscr{F}_x$ can be rewritten in the following form:

$$\Pr(\phi_i | \mathscr{M}_m) = \frac{1}{\sqrt{(2\pi)^{d_\phi} |\mathbf{H}_m|}} e^{-\frac{1}{2} \tilde{\phi}_{mi}^T \mathbf{H}_m^{-1} \tilde{\phi}_{mi}}, \tag{4.14}$$

where $\tilde{\phi}_{mi} = \phi_i - \mu_m$. It is worth noting that the neighborhood model could not have been computed in the input space X so far. We will introduce its solution method in the next section.

Due to the complicated data distribution and the nonlinear mapping between the textual and visual spaces, a single mapping model is insufficient in modeling the relationship between different media. To this end, we develop a probabilistic mixture model to characterize the cross-media mapping. Given the model, a log-likelihood function is defined based on the joint probability of N cross-media data pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ as follows:

$$\begin{aligned} \mathscr{L}_f &= \ln \prod_{i=1}^{N} \Pr(\phi_i, \psi_i) \\ &= \ln \prod_{i=1}^{N} \sum_{m=1}^{M} \omega_m \Pr(\phi_i | \mathscr{M}_m) \Pr(\psi_i | \phi_i, \mathscr{M}_m), \end{aligned} \tag{4.15}$$

where M is the number of components in the mixture model, and ω_m is the weight of the m-th component $\mathscr{M}_m$ with $\sum_{m=1}^{M} \omega_m = 1$ and $\omega_m \geq 0$. In the mixture model, the first probabilistic term aims to make close points share the same component $\mathscr{M}_m$, and the second term focuses on modeling the relationship between two modalities.

4.3.3. CONSTRAINTS IN THE MODEL

According to Eq.4.13, $\mathbf{A}_m$ and $\mathbf{B}_m$ are the coefficient matrices used to reconstruct $\mathbf{U}_m$ and $\mathbf{V}_m$ based on the kernel matrices $\mathbf{K}_y$ and $\mathbf{K}_x$, respectively. Here, we consider two extra constraints.

Smoothness

In general, two close data points in the input spaces X and Y are expected to have images close together in the latent semantic space S. To this end, we introduce a smoothness constraint that is defined as follows:

$$\begin{aligned} J_A &= \sum_{i\sim j} \|\mathbf{A}(K_{y,i} - K_{y,j})\|_2^2 \\ &\leq \|\mathbf{P}\|_F^2 \|\mathbf{A}\|_F^2 \ , \\ J_B &= \sum_{i\sim j} \|\mathbf{B}(K_{x,i} - K_{x,j})\|_2^2 \\ &\leq \|\mathbf{Q}\|_F^2 \|\mathbf{B}\|_F^2 \ , \end{aligned} \tag{4.16}$$

where $i \sim j$ denotes that the i-th and j-th data points are close, and $\mathbf{P}$ and $\mathbf{Q}$ are the matrices whose columns are the vectors $\{(K_{y,i} - K_{y,j})\}$ and $\{(K_{x,i} - K_{x,j})\}$, respectively, in a certain order for all $i \sim j$. Thus, we characterize the smoothness of the cross-media mapping based on Eq.4.16 as follows:

$$J_{sm}(\mathbf{A}_m, \mathbf{B}_m) = \lambda_{A,m}\|\mathbf{A}_m\|_F^2 + \lambda_{B,m}\|\mathbf{B}_m\|_F^2 \ , \tag{4.17}$$

where for simplicity, we use parameters $\lambda_{A,m}$ and $\lambda_{B,m}$ to replace the exact Frobenius norm of $\mathbf{P}$ and $\mathbf{Q}$, respectively. In our work, we let $\lambda_{A,m} = \lambda_{B,m} = 1$ and use a single λ_1 to control the importance of the smoothness term. To obtain a smooth mapping model, $J_{sm}(\mathbf{A}, \mathbf{B})$ needs to be constrained to a small value.

Sparseness

The rows of $\mathbf{U}_m$ and $\mathbf{V}_m$ can be considered as a new basis (possibly nonorthonormal) for the projection of examples $\{\psi_i\}$ and $\{\phi_i\}$, respectively, and can be linearly reconstructed by these examples. To make each basis tend to represent some specific semantics held by a subset of examples, we expect to reconstruct the rows of $\mathbf{U}_m$ and $\mathbf{V}_m$ using a few examples by enforcing each row of $\mathbf{A}_m$ and $\mathbf{B}_m$ to have a few non-zero elements. We call the characteristics sparseness and formulate it using the L_1-norm as follows:

$$J_{sp}(\mathbf{A}_m, \mathbf{B}_m) = \|\mathbf{A}_m\|_1 + \|\mathbf{B}_m\|_1 \ . \tag{4.18}$$

Incorporating both constraints into our problem, we have the final optimization problem to compute cross-media correlation:

$$\max_{\Theta} \mathscr{L}_f - \sum_{m=1}^{M} \left(\lambda_1 J_{sm}(\mathbf{A}_m, \mathbf{B}_m) + \lambda_2 J_{sp}(\mathbf{A}_m, \mathbf{B}_m)\right) \ , \tag{4.19}$$

where $\Theta = \{\omega_m, \mu_m, \mathbf{H}_m, \mathbf{A}_m, \mathbf{B}_m, \Sigma_m\}_{m=1}^{M}$ is the parameter set, and λ_1 and λ_2 are used to control the balance between the terms.

4.4. OPTIMIZATION, ALGORITHM AND ANALYSIS

4.4.1. OPTIMIZATION AND ALGORITHM

Similar to Wang et al.'s work [135], we define the following notations as shown in Table 4.1. The first five rows in this table formulate the traditional estimation of Gaussian mixture models in the input space based on EM [136]. Here, the superscript (t) refers to the t-th iteration.

Table 4.1: Notations.

$p_{mi}^{(t)} = \Pr(\mathscr{M}_m \vert \phi_i, \psi_i, \Theta^{(t)})$
$w_{mi}^{(t)} = \sqrt{p_{mi}^{(t)} / \sum_{j=1}^{N} p_{mj}^{(t)}}$
$\mu_m^{(t)} = \sum_{i=1}^{N} (w_{mi}^{(t)})^2 \phi_i$
$\tilde{\phi}_{mi}^{(t)} = \phi_i - \mu_m^{(t)}$
$\mathbf{H}_m^{(t)} = \sum_{i=1}^{N} (w_{mi}^{(t)})^2 \tilde{\phi}_{mi} \tilde{\phi}_{mi}^T$
$(\mathbf{K}_x)_{ij} = \phi_i \cdot \phi_j = f_\phi(\mathbf{x}_i, \mathbf{x}_j)$
$(\mathbf{K}_{x,m})_{ij}^{(t)} = w_{mi}^{(t)} \phi_i \cdot w_{mj}^{(t)} \phi_j$
$(\tilde{\mathbf{K}}_{x,m})_{ij}^{(t)} = w_{mi}^{(t)} \tilde{\phi}_{mi} \cdot w_{mj}^{(t)} \tilde{\phi}_{mj}$
$(\mathbf{K}'_{x,m})_{ij}^{(t)} = \phi_i \cdot w_{mj}^{(t)} \phi_j$
$(\tilde{\mathbf{K}}'_{x,m})_{ij}^{(t)} = \tilde{\phi}_{mi} \cdot w_{mj}^{(t)} \tilde{\phi}_{mj}$

The optimization problem Eq. 4.19 is different from previous regularization-based learning problems because it contains the hidden information. More specifically, we do not know which component $\mathscr{M}_m$ "generates" each pair $(\mathbf{x}_i, \mathbf{y}_i)$. To solve the optimization problem, we present a hybrid optimization algorithm based on EM and subgradient ascent. The parameters of the proposed model are $\Theta = \{\omega_m, \mu_m, \mathbf{H}_m, \mathbf{A}_m, \mathbf{B}_m, \Sigma_m\}_{m=1}^{M}$, where the first three parameters describe the neighborhood model, and the rest are for cross-media mapping.

Based on the EM algorithm, we define the following function $\tilde{\mathscr{L}}_f$ in the expectation step to help optimize problem Eq. 4.19.

$$\begin{aligned} \tilde{\mathscr{L}}_f &= \sum_{m=1}^{M} \sum_{i=1}^{N} p_{mi} \ln \left(\omega_m \Pr(\phi_i \vert \mathscr{M}_m) \Pr(\psi_i \vert \phi_i, \mathscr{M}_m) \right) \\ &= \sum_{m=1}^{M} \sum_{i=1}^{N} p_{mi} \left(\ln \omega_m + \ln \Pr(\phi_i \vert \mathscr{M}_m) + \ln \Pr(\psi_i \vert \phi_i, \mathscr{M}_m) \right). \end{aligned} \tag{4.20}$$

According to EM, the growth of $\tilde{\mathscr{L}}_f$ can increase $\mathscr{L}_f$ shown in problem Eq. 4.19.

By setting the partial derivative of $\tilde{\mathscr{L}}_f$ to zero, we can easily achieve

$$\omega_m^{(t)} = \frac{1}{N}\sum_{i=1}^{N} p_{mi}^{(t)} \ , \tag{4.21}$$

where $p_{mi}^{(t)}$ is defined in Table 4.1 and can be expanded as:

$$p_{mi}^{(t)} = \frac{\omega_m^{(t-1)} \Pr(\phi_i|\mathscr{M}_m, \Theta^{(t-1)}) \Pr(\psi_i|\phi_i, \mathscr{M}_m, \Theta^{(t-1)})}{\sum_{k=1}^{M} \omega_k^{(t-1)} \Pr(\phi_i|\mathscr{M}_k, \Theta^{(t-1)}) \Pr(\psi_i|\phi_i, \mathscr{M}_k, \Theta^{(t-1)})}. \tag{4.22}$$

The feature space $\mathscr{F}_x$ is usually of high dimension and cannot be represented explicitly. Hence, we do not directly compute the distribution $\Pr(\phi_i|\mathscr{M}_m, \Theta^{(t-1)})$ in Eq.4.15 and Eq.4.21 (sometimes $\Theta^{(t-1)}$ may be omitted to save space) and estimate the parameters $\{\mu_m, \mathbf{H}_m\}_{m=1}^{M}$ in the feature space. Instead, we may estimate the distribution in the input space with the kernel trick. First, based on the work in [137], we can rewrite the exponent term in Eq.4.14 as:

$$\begin{aligned} \tilde{\phi}_{mi}^T \mathbf{H}_m^{-1} \tilde{\phi}_{mi} =& \tilde{\phi}_{mi}^T \mathbf{V} \Lambda^{-1} \mathbf{V}^{\mathrm{T}} \tilde{\phi}_{mi} \\ =& \sum_{j=1}^{d_\phi} y_j^2 / \lambda_j \ , \end{aligned} \tag{4.23}$$

where $\mathbf{V}$ and $\Lambda^{-1} = diag(\lambda_1^{-1}, \lambda_2^{-1}, \cdots, \lambda_{d_\phi}^{-1})$ denote the matrices of the eigenvectors and eigenvalues of $\mathbf{H}_m^{-1}$, respectively, and $y_j = \tilde{\phi}_{mi}^T \mathbf{V}_j$ is the projection of $\tilde{\phi}_{mi}^T$ over the j-th eigenvector $\mathbf{V}_j$. We note that $\tilde{\mathbf{K}}_{x,m}$ and the bandwidth matrix $\mathbf{H}_m$ have the same nonzero eigenvalues $\{\lambda_j\}$. It was proved in [135] that

$$y_j = \beta_j^T \Gamma_{\cdot,i} \ , \tag{4.24}$$

where β_j is the eigenvector of $\tilde{\mathbf{K}}_{x,m}$ corresponding to the eigenvalue λ_j, and $\Gamma_{\cdot,i}$ is the column of $\tilde{\mathbf{K}}'_{x,m}$ corresponding to $\mathbf{x}_i$. Note that d_ϕ is unknown due to the implicit feature map ϕ, and we approximately estimate the distribution as the marginal density function by keeping d'_ϕ ($d'_\phi < d_\phi$) principal components that correspond to the d'_ϕ largest nonzero eigenvalues and discard the rest in Eq.4.23. Moreover, the factor $(2\pi)^{d_\phi/2}\mathbf{H}_m^{1/2}$ in Eq.4.14 can be replaced by $(2\pi)^{d'_\phi/2}\prod_{j=1}^{d'_\phi}\lambda_j^{1/2}$. Then, we can iteratively estimate $\Pr(\phi_i|\mathscr{M}_m)$ in Eq.4.20 by updating the kernel-matrix parameters $\tilde{\mathbf{K}}_{x,m}$ and $\tilde{\mathbf{K}}'_{x,m}$ shown in Table 4.1 in the input space, instead of μ_m and $\mathbf{H}_m$ in the feature space, since both sets of parameters describe the same distribution. Note that $\tilde{\mathbf{K}}_{x,m}$ and $\tilde{\mathbf{K}}'_{x,m}$ can be easily computed as the centralized versions of $\mathbf{K}_{x,m}$ and $\mathbf{K}'_{x,m}$, respectively.

For the update of the parameters $\{\mathbf{A}_m, \mathbf{B}_m, \Sigma_m | m = 1, 2, \cdots, M\}$, we build a new optimization function based on problem Eq. 4.19 and Eq.4.20:

$$\mathscr{Q} = \sum_{m=1}^{M} \left(\sum_{i}^{N} p_{mi} \log \Pr(\psi_i|\phi_i, \mathscr{M}_m) - J(\mathbf{A}_m, \mathbf{B}_m) \right), \tag{4.25}$$

where $J(\cdot,\cdot) = \lambda_1 J_{sm}(\cdot,\cdot) + \lambda_2 J_{sp}(\cdot,\cdot)$. $\mathscr{Q}$ includes the non-differentiable terms of $\|\cdot\|_1$ for the sparseness constraint, and thus a closed-form solution cannot be obtained by directly taking the derivative. We use the subgradient ascent scheme to iteratively maximize $\mathscr{Q}$. At each time t, we compute the subgradients as

$$\nabla_{\Sigma_m}\mathscr{Q} = -\frac{1}{2}\sum_i p_{mi}(\Sigma_m^{-1} - D_{m,i}\Sigma_m^{-2}D_{m,i}^T) ,$$
$$\nabla_{\mathbf{A}_m}\mathscr{Q} = -\sum_i p_{mi}\Sigma_m^{-1}\mathbf{A}_m D_{m,i}^T - \lambda_1\mathbf{A}_m - \lambda_2\Delta_{\mathbf{A}_m} ,$$
$$\nabla_{\mathbf{B}_m}\mathscr{Q} = \sum_i p_{mi}\Sigma_m^{-1}\mathbf{B}_m D_{m,i}^T - \lambda_1\mathbf{B}_m - \lambda_2\Delta_{\mathbf{B}_m} ,$$

where $D_{m,i} = \mathbf{A}_m K_{x,i} - \mathbf{B}_m K_{y,i}$, and Δ is defined as

$$(\Delta_{\mathbf{A}_m})_{ij} = \mathrm{sgn}((\mathbf{A}_m)_{ij}),\ (\Delta_{\mathbf{B}_m})_{ij} = \mathrm{sgn}((\mathbf{B}_m)_{ij}) .$$

Here, sgn(z) outputs 1 when $z > 0$, 0 when $z < 0$, and a random value uniformly distributed in [-1,1] when $z = 0$. Given the subgradients, we update the solution for Σ_m, $\mathbf{A}_m$ and $\mathbf{B}_m$ to maximize $\mathscr{Q}$ as follows:

$$\begin{aligned}\Sigma_m^{(t+1)} &= \Sigma_m^{(t)} + \eta_{\Sigma_m}^{(t)} \cdot \nabla_{\Sigma_m}\mathscr{Q} ,\\ \mathbf{A}_m^{(t+1)} &= \mathbf{A}_m^{(t)} + \eta_{\mathbf{A}_m}^{(t)} \cdot \nabla_{\mathbf{A}_m}\mathscr{Q} ,\\ \mathbf{B}_m^{(t+1)} &= \mathbf{B}_m^{(t)} + \eta_{\mathbf{B}_m}^{(t)} \cdot \nabla_{\mathbf{B}_m}\mathscr{Q} ,\end{aligned} \tag{4.26}$$

where $\eta_{\Sigma_m}^{(t)}$, $\eta_{\mathbf{A}_m}^{(t)}$ and $\eta_{\mathbf{B}_m}^{(t)}$ are the step sizes at time t. In the experiment, we set the step size to $1/t$.

Eigenvalue decomposition of large-scale matrices

When an example $\mathbf{x}_i$ is far from the center of the Gaussian component $\mathscr{M}_m$, it belongs to this component with low probability p_{mi}. Hence, we can set the corresponding columns and rows of $\tilde{\mathbf{K}}_{x,m}$ to zero to obtain an approximation, and perform eigenvalue decomposition on a smaller matrix after the elementary transformation of the matrices. Let $p_m = \max_j p_{mj}$, and in the experiments, we set the i-th columns and rows of $\tilde{\mathbf{K}}_{x,m}$ to zero if $p_{mi} < 0.01 p_m$.

Parameter initialization

First, we use the K-means clustering algorithm with the training data $\{\mathbf{x}_i\}_{i=1}^N$ and achieve M clusters in the input space. Then, we compute the values of p_{mi}, $\mathbf{K}_{x,m}$, $\tilde{\mathbf{K}}_{x,m}$, $\mathbf{K}'_{x,m}$ and $\tilde{\mathbf{K}}'_{x,m}$ over the hard partitions as the parameters at the time $t = 0$, $\Sigma_m^{(0)}$, $\mathbf{A}_m^{(0)}$ and $\mathbf{B}_m^{(0)}$ are set randomly.

4.4.2. CONVERGENCE ANALYSIS

Our model includes a hidden variable, i.e., $\mathscr{M}_m$, to indicate which local model that a pair of data points follows. The EM algorithm is a powerful tool for solving

Algorithm 2 KMM parameter estimation algorithm

Require: Image-text paired documents $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, each including an image and a textual document with the same semantics, kernel functions k_ϕ and k_ψ, parameters M, λ_1, λ_2, d_s, step sizes $\eta_{\Sigma_m}^{(t)}$, $\eta_{\mathbf{A}_m}^{(t)}$ and $\eta_{\mathbf{B}_m}^{(t)}$, and the maximum number of iterations T;

Ensure: The estimated parameter set $\Theta = \{\omega_m, \tilde{\mathbf{K}}_{x,m}, \tilde{\mathbf{K}}'_{x,m}, \mathbf{A}_m, \mathbf{B}_m, \Sigma_m\}_{m=1}^{M}$;

1: Initialize parameter set $\Theta^{(0)} = \{\omega_m^{(0)}, \tilde{\mathbf{K}}_{x,m}^{(0)}, \tilde{\mathbf{K}}_{x,m}^{\prime(0)}, \mathbf{A}_m^{(0)}, \mathbf{B}_m^{(0)}, \Sigma_m^{(0)}\}_{m=1}^{M}$, $t = 0$;
2: **repeat**
3: $t = t + 1$, and step sizes $\eta_{\Sigma_m}^{(t)}, \eta_{\mathbf{A}_m}^{(t)}, \eta_{\mathbf{B}_m}^{(t)} = 1/t$;
4: Compute $p_{mi}^{(t)}$ with Eq.4.22 based on $\Theta^{(t-1)}$;
5: Update $\omega_m^{(t)}$ with Eq.4.21, $\tilde{\mathbf{K}}_{x,m}^{(t)}$ and $\tilde{\mathbf{K}}_{x,m}^{\prime(t)}$ based on Table 4.1, and then get $\tilde{\Theta}^{(t)}$;
6: Update $\Sigma_m^{(t)}$, $\mathbf{A}_m^{(t)}$ and $\mathbf{B}_m^{(t)}$ with Eq.4.26, and then get $\Theta^{(t+1)}$;
7: **until** $t \geq T$.

such problems and can generally guarantee that the iterative optimization converges to a local optimal solution. In our work, we present a hybrid optimization algorithm based on the combination of EM and subgradient ascent. In this subsection, we introduce two notations: X denotes the observed data and z hidden states commonly used in the EM algorithm, which correspond to $\{(\phi_i, \psi_i)\}$ and $\{\mathscr{M}_m\}$, respectively, in our model. Here, we rewrite $\mathscr{L}_f = \ln \prod_{i=1}^{N} \Pr(\phi_i, \psi_i)$ in Eq.4.15, i.e., $\ln \Pr(X|\Theta)$, as $\mathscr{L}_f(\Theta^{(t)})$ to emphasize the parameters at a particular time t, and we define the following variable:

$$\begin{aligned}\ell_f(\Theta|\Theta^{(t)}) =& \mathscr{L}_f(\Theta^{(t)}) + \sum_z \Pr(z|X,\Theta^{(t)}) \ln\left(\frac{\Pr(X,z|\Theta)}{\Pr(X,z|\Theta^{(t)})}\right) \\ =& \mathscr{L}_f(\Theta^{(t)}) + l_1(\Theta|\Theta^{(t)}) + l_2(\Theta|\Theta^{(t)}) \\ & - \sum_z \Pr(z|X,\Theta^{(t)}) \ln \Pr(X,z|\Theta^{(t)}) \ , \end{aligned} \tag{4.27}$$

where

$$\begin{aligned} l_1(\Theta|\Theta^{(t)}) &= \sum_z \Pr(z|X,\Theta^{(t)}) \ln P_1(X,z|\Theta) \ , \\ l_2(\Theta|\Theta^{(t)}) &= \sum_z \Pr(z|X,\Theta^{(t)}) \ln P_2(X,z|\Theta) \ , \end{aligned}$$

and $\Theta^{(t)}$ denotes the current parameters at time t. In this subsection, $P_1(\cdot)$ and $P_2(\cdot)$ correspond to $\omega_m \Pr(\phi_i|\mathscr{M}_m)$ and $\Pr(\psi_i|\phi_i, \mathscr{M}_m)$, respectively, in Eq.4.16. Based on Jensens inequality, we have that $\mathscr{L}_f(\Theta) \geq \ell_f(\Theta|\Theta^{(t)})$, and then that

$$\mathscr{L}_f(\Theta) - J(\Theta) \geq \ell_f(\Theta|\Theta^{(t)}) - J(\Theta) \ . \tag{4.28}$$

where $J(\Theta)$ represents the regularization terms in problem Eq. 4.19 and Eq.4.25, i.e., $\sum_{m=1}^{M} J(\mathbf{A}_m, \mathbf{B}_m)$. In Eq.4.28, the equality holds if and only if $\Theta = \Theta^{(t)}$.

In the optimization process shown in Algorithm 1, we have the following two steps in each iteration. 1) We update ω_m, $\tilde{\mathbf{K}}_{x,m}$ and $\tilde{\mathbf{K}}'_{x,m}$ based on the EM algorithm to maximize $\sum_{m,i} p_{mi}\big(\ln\omega_m + \ln \Pr(\phi_i|\mathscr{M}_m)\big)$ in Eq.4.21, i.e., $l_1(\Theta|\Theta^{(t)})$, and obtain the parameter denoted by $\tilde{\Theta}^{(t)}$; 2) By fixing the updated parameters in step 1, we update parameters Σ_m, $\mathbf{A}_m$ and $\mathbf{B}_m$ via Eq.4.26 to maximize $\mathscr{Q}$, i.e., $l_2(\Theta|\Theta^{(t)}) - J(\Theta)$. Consequently, based on the above steps, we increase the right side of Eq.4.28 and obtain the updated parameters, denoted by $\Theta^{(t+1)}$. According to Eq.4.28, $\mathscr{L}_f(\Theta^{(t+1)}) - J(\Theta^{(t+1)}) > \mathscr{L}_f(\Theta^{(t)}) - J(\Theta^{(t)})$. Consequently, our algorithm will converge to an (local) optimal solution. Fig. 4.2(b) illustrates the optimization process.

4

4.4.3. COMPLEXITY ANALYSIS

The computational complexity of parameter estimation is mainly derived from the update of kernel matrices, e.g., $\tilde{\mathbf{K}}_{x,m}$, the eigenvalue decomposition of $\tilde{\mathbf{K}}_{x,m}$, the subgradient computation and the update in Eq.4.26. Suppose the numbers of examples handled by each component $\mathscr{M}_m$, denoted by N_m, are the same and do not change as the amount of data increases; the proposed algorithm includes the following four main parts: 1) computing $\tilde{\mathbf{K}}_{x,m}$ in $O(N_m^2 M)$ time, 2) performing the eigenvalue decomposition of $\tilde{\mathbf{K}}_{x,m}$ in $O(N_m^3 M)$ time, 3) computing the subgradients with respect to Σ_m and $\mathbf{A}_m$ ($\mathbf{B}_m$) in $O(d_s^3 M + d_s N_m^2 M)$ time and $O(d_s^3 M + d_s^2 NM + d_s N_m^2 M + d_s NM)$ time, respectively, and 4) updating Σ_m and $\mathbf{A}_m$ ($\mathbf{B}_m$) in $O(d_s^2 M)$ time and $O(d_s N)$ time, respectively. Suppose the algorithm converges in T iterations; the total computational complexity is $O(NMT)$ by keeping the higher-order terms in the above analysis.

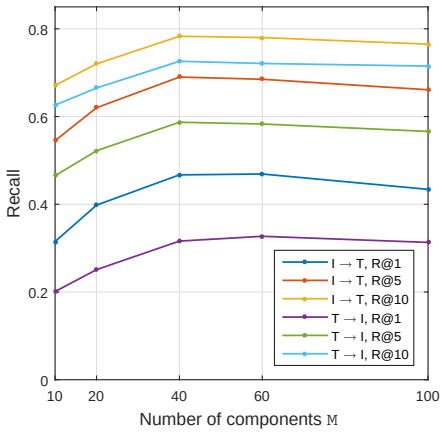
4.5. EXPERIMENTAL RESULTS

4.5.1. DATASET AND EXPERIMENTAL SETTING

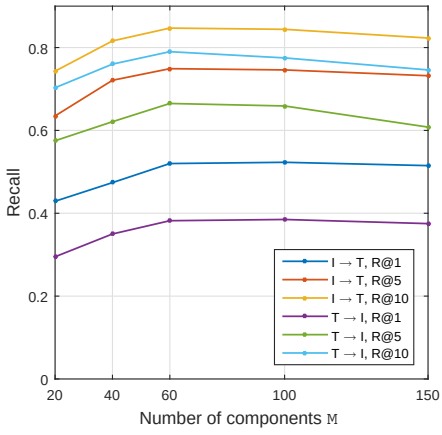
Four public real-world datasets are used in our experiments.

- *Flickr8K* dataset [138] consists of 8,000 images from the Flickr.com website, which focuses on people or animals performing actions. For each image, five captions were generated by different annotators using a crowdsourcing service. The dataset is split into disjoint training, validation, and test sets with 6,000, 1,000, and 1,000 pairs, respectively.

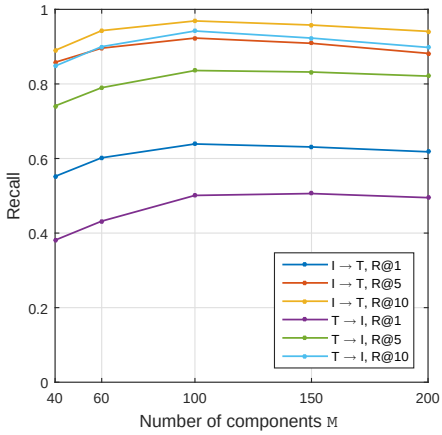
- *Flickr30K* dataset [24] extends the Flicker8K and consists of 31,783 images of everyday activities, events and scenes, each paired with five captions, i.e., a total of 158,915 captions. The captions were annotated in a similar style as in Flicker8K. We use 1,000 examples for testing, 1,000 examples for validation and the rest for training.



(a) Flickr8K



(b) Flickr30K



(c) MSCOCO

Figure 4.3: The performance of cross-media retrieval versus the number of components M in the case of "1 image vs. 1 caption" on three datasets. "I → T" and "T → I" denote "image → text" and "text → image", respectively.

- *MSCOCO* dataset [21] contains 123,287 images, each corresponding to 5 captions. Similar to [139], we randomly generate the splits that contain 5,000 images with corresponding captions for both validation and testing, and the rest of the images are used for training. The results are reported on a subset of 1,000 testing images.

- *NUS-WIDE-10K* dataset [140] has 10,000 image/text pairs in total, selected evenly from the 10 largest categories of the NUS-WIDE dataset. The dataset is split into three subsets following [130]: training set with 8,000 pairs, testing set with 1,000 pairs and validation set with 1,000 pairs.

We implement 5 independent experiments to alleviate the variation caused by random splits of datasets.

The data are represented as follows.

- *Image representation*: In the experiments, we employ two pre-trained deep networks on ImageNet, i.e., VGG-16 networks [4] and ResNet-152 [5]. We use the images resized to 224×224 as the input for both networks, and achieve 4096-dimensional feature vectors from VGG and 2048-dimensional vectors from ResNet.

- *Text representation*: We extract textual features based on Word2vec [141]. We represent every word in a commonly used 150-dimensional embedding space, and then cluster them into K groups. Finally, we employ a bag-of-words representation to describe an text instance based on the feature vectors of words. In the experiments, we let $K = 500$.

Cross-media retrieval includes two tasks: text retrieval given a query of an image and image retrieval given a textual query, which are denoted by "image → text" and "text → image", respectively. We evaluate the performance with R@r that denotes the recall at r for both tasks. Since Flickr8k, Flickr30k and MSCOCO contain 5 captions per image, we evaluate the proposed approach in two cases: 1) "1 image vs. 1 caption", in which each caption is considered as a response or a query in the retrieval, and the recall at r for "image → text" task is computed based on whether at least one of the correct captions is among the first r retrieved ones [142], and 2) "1 image vs. 5 captions", in which the 5 captions corresponding to an image are concatenated as a response or a query [128]. In the cases of "1 image vs. 1 caption" and "1 image vs. 5 captions", we train KMM based on the pair of an image and each of its 5 captions [103] and the pair of an image and its concatenated captions [128], respectively. Regarding NUS-WIDE-10K, like [130], we consider the set of multiple tags for an image as a text instance in both retrieval tasks and evaluate the performance with the mean average precision (mAP) score.

In the experiment, we compare the proposed approach with the following state-of-the-art methods.

- Deep CCA [128]: representing images and captions using deep neural networks and then correlating them by CCA.

- HGLMM and GMM+HGLMM [139]: combining Gaussian and Laplacian distributions into one hybrid distribution model that can benefit from the properties of the two distributions.
- MLLM [43]: a mixture of local linear mapping model with VGG-16-based visual representation and Word2vec-based text representation.
- 2-Way Net [102]: employing two tied neural network channels that project the two views into a common, maximally correlated space using Euclidean loss.
- Embedding Networks [103]: learning a shared latent embedding space based on two-way networks with a maximum-margin ranking loss and neighborhood constraints.
- DVSA [143]: an alignment model based on the combination of CNNs over image regions and bidirectional recurrent neural networks over sentences.
- OrderEmbedding [144]: learning the embeddings of images and captions by defining a loss function that encourages the order-violation penalty for ground truth caption-image pairs to be lower than that for all other pairs, by a margin.
- DSvEL [142]: a new two-path neural network with a visual path that leverages recent space-aware pooling mechanisms.
- CSE [145]: using CNNs to represent images and sentences and combining mid-level representations and global semantic learning.
- CCL [130]: fusing multi-grained features and learning the correlation based on the constraints of the intra-modality semantic category and the inter-modality pairwise similarity.
- RRF-Net [146]: a model that adapts the recurrent mechanism to residual learning and integrates the intermediate recurrent outputs.

4.5.2. PARAMETER TUNING AND ANALYSIS

The key parameters of KMM include the number of components M in Eq.4.15, the balance control parameters λ_1 and λ_2 in problem Eq. 4.19, and the dimension, d_s of semantic space S. To maximize the performance over validation sets, we determine the parameters by searching on the following grids: $\lambda_1, \lambda_2 \in \{10^2, 10^1 \cdots, 10^{-3}\}$, $M \in \{10, 20, 40, 60, 100, 150, 200\}$, and $d_s \in \{20, 50, 100, 150, 200\}$. In the experiment, we set $d_s = 50$ for Flickr8K, Flickr30K and NUS-WIDE-10K, and $d_s = 100$ for MSCOCO. To avoid inner product computation in the implicit feature spaces $\mathscr{F}_x$ and $\mathscr{F}_y$, we introduce the kernel function in Section 4.3.2 to achieve the computational results in the input spaces. We choose a polynomial kernel of degree 2, i.e., $f_\phi(\mathbf{x}_k, \mathbf{x}_l) = (\mathbf{x}_k \cdot \mathbf{x}_l + 1)^2$, $f_\psi(\mathbf{y}_k, \mathbf{y}_l) = (\mathbf{y}_k \cdot \mathbf{y}_l + 1)^2$, via experimentation.

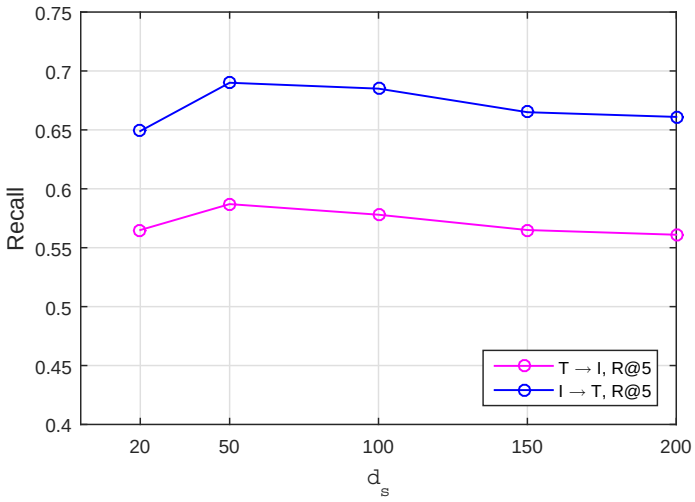


Figure 4.4: The effect of the dimension, d_s, of semantic space on the retrieval performance (R@5) of KMM with $M = 40$ in the case of "1 image vs 1 caption" on Flickr8K. "I → T" and "T → I" denote "image → text" and "text → image", respectively.

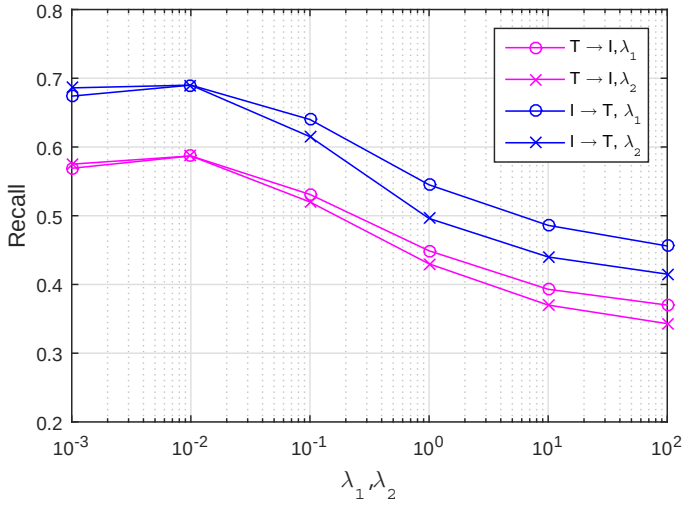


Figure 4.5: The effect of λ_1 and λ_2 on the retrieval performance (R@5) of KMM with $M = 40$ in the case of "1 image vs 1 caption" on Flickr8K. "I → T" and "T → I" denote "image → text" and "text → image", respectively. We change one parameter by setting the other to the optimal value, i.e., 10^{-2}.

Table 4.2: Performance (percent) comparison of bi-directional retrieval on Flickr8K. The top two parts in the body of the table correspond to the case of "1 image vs. 1 caption" and the bottom two parts correspond to the case of "1 image vs. 5 captions".

Approaches	Visual Backend	Flickr8K					
		Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10
HGLMM [139]	VGG	28.5	58.4	71.7	20.6	49.4	64.0
GMM+HGLMM [139]	VGG	31.0	59.3	73.7	21.2	50.0	64.8
2-Way Net [102]	VGG	43.4	63.2	-	29.3	49.7	-
MLLM [43]	VGG	34.2	59.9	70.8	26.9	54.0	67.9
KMM (without sparseness)	ResNet	45.7	68.5	77.3	31.0	57.1	71.2
KMM (without smoothness)	ResNet	44.6	67.2	76.3	30.1	55.8	70.6
KMM	VGG	43.4	65.9	74.5	29.0	56.1	69.3
KMM	ResNet	**46.7**	**69.0**	**78.3**	**31.6**	**58.7**	**72.6**
Deep CCA [128] (1 ima. vs. 5 cap.)	AlexNet	28.2	56.1	69.8	26.3	54.0	67.5
MLLM [43] (1 ima. vs. 5 cap.)	VGG	32.5	59.2	70.3	27.8	54.7	68.9
KMM (1 ima. vs. 5 cap.)	VGG	42.9	65.6	74.6	29.8	56.3	70.2
KMM (1 ima. vs. 5 cap.)	ResNet	**46.1**	**68.9**	**78.5**	**32.5**	**59.1**	**73.3**

Fig. 4.3 illustrates the effect of parameter M on the performance of cross-media retrieval for three datasets in the case of "1 image vs. 1 caption". On the whole, we observe that the recalls reach the highest values at $M = 40$ and 60 for Flickr8K and Flickr30K, respectively. For the more complex MSCOCO dataset, a larger value, $M = 100$, can produce better performance than the other values of M. The phenomenon is consistent with our intuition. That is, a model of larger capacity, i.e., the one with larger M in this work, is required for modeling a more complex dataset. In addition, the performances measured by different metrics on a specific dataset likely do not reach the highest value at the same M. For example, for Flickr8K, the recall R@1 in task "text→ image" is 31.6% at $M = 40$, which is slightly lower than 32.7% at $M = 60$. We also notice that a value of M that is too large may decrease the size of the local region $\mathscr{R}_m$ that supports local model $\mathscr{M}_m$, which tends to cause over-fitting in the learning of parameters $\mathbf{A}_m$, $\mathbf{B}_m$ and Σ_m and affects the performance of cross-media retrieval.

Fig. 4.4 shows the effect of the dimension d_s of semantic spaces on the retrieval performance (R@5) of KMM with $M = 40$ in the case of "1 image vs. 1 caption" on Flickr8K. As seen, the dimension of semantic spaces has effects on the performance. More specifically, the recall reaches the peak at $d_s = 50$ and then begins to degrade. An appropriate dimension for a latent semantic space depends on the complexity of semantics contained in datasets. A lower dimensional semantic space may result in an insufficient capacity to represent the distribution of semantics, while a higher dimension may cause a looser distribution of semantics as well as larger sizes of transformation matrices $\mathbf{U}_m$ and $\mathbf{V}_m$.

Fig. 4.5 illustrates the effect of parameters λ_1 and λ_2 on the retrieval performance (R@5) of KMM with $M = 40$ in the case of "1 image vs. 1 caption" on Flickr8K. In the figure, we show the effect of one parameter while setting the other to the optimal value. By experiments, we find that the retrieval performance peaks at $\lambda_1 = \lambda_2 = 10^{-2}$ and then retrieval performance begins to degrade as λ_1 or λ_2 continues to be added. In general, we find λ_1 leads a faster increase and slower degradation of performance than λ_2 as parameters are added. The results indicate that the smoothness term plays a more important role than the sparseness term in maintaining a good retrieval performance. In the experiments, we set λ_1 or λ_2 to 10^{-2} for all datasets. We conduct a further analysis for λ_1 and λ_2 by an ablation study in the next subsection.

4.5.3. PERFORMANCE ON CROSS-MEDIA RETRIEVAL

Ablation study

To further reveal the contribution of the two constraints in problem Eq. 4.19, we test the performance of KMM with three configurations. The variants include: 1) KMM (without smoothness), which is obtained by removing the smoothness term, 2) KMM (without sparseness), which ignores the L1-norm regularization that constrains the sparseness of learning results, and 3) KMM, which is the full version formulated in problem Eq. 4.19. Table 4.2, Table 4.3 and Table 4.4 show

Table 4.3: Performance (percent) comparison of bi-directional retrieval on Flickr30K. The top two parts in the body of the table correspond to the case of "1 image vs. 1 caption" and the bottom two parts correspond to the case of "1 image vs. 5 captions".

Approaches	Visual Backend	Flickr30K					
		Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10
HGLMM [139]	VGG	34.4	61.0	72.3	24.4	52.1	65.6
GMM+HGLMM [139]	VGG	35.0	62.0	73.8	25.0	52.7	66.0
2-Way Net [102]	VGG	49.8	67.5	-	36.0	55.6	-
MLLM [43]	VGG	44.5	62.8	73.4	28.1	53.6	65.8
DSvEL [142]	ResNet	46.5	72.0	82.2	34.9	62.4	73.5
CSE [145]	ResNet	44.6	74.3	83.8	36.9	**69.1**	**79.6**
Embedding Networks [103]	VGG	43.2	71.6	79.8	31.7	61.3	72.4
KMM (without sparseness)	ResNet	50.9	74.0	84.2	38.0	64.9	77.4
KMM (without smoothness)	ResNet	50.3	71.6	83.4	37.2	63.9	76.6
KMM	VGG	49.1	72.2	81.5	35.1	60.1	72.8
KMM	ResNet	**52.0**	**74.9**	**84.7**	**38.2**	66.5	79.0
Deep CCA [128] (1 ima. vs. 5 cap.)	AlexNet	27.9	56.9	68.2	26.8	52.9	66.9
MLLM [43] (1 ima. vs. 5 cap.)	VGG	43.7	62.2	73.5	28.5	54.1	66.1
CCL [130] (1 ima. vs. 5 cap.)	VGG	37.7	69.4	81.1	37.3	**68.4**	**80.0**
KMM (1 ima. vs. 5 cap.)	VGG	48.5	71.9	81.2	36.2	60.9	73.2
KMM (1 ima. vs. 5 cap.)	ResNet	**51.6**	**75.4**	**85.3**	**39.4**	66.9	79.5

the comparison results of the variants in the case of "1 image vs. 1 caption" on Flickr8K, Flickr30K and MSCOCO. From the tables, we observe that the variants KMM (without smoothness) and KMM (without sparseness) generally perform slightly worse than KMM (with visual representation using ResNet). Both variants have a degradation of 1.0 ~ 3.3% on the whole compared with KMM. From the figures, we observe that the smoothness term plays a more important role than the sparseness term in improving performance. The main cause is that smoothness may enforce two similar examples to be close together in the latent space.

Performance comparison

First, we evaluate and analyze the performance of the proposed approach in the case of "1 image vs. 1 caption" (i.e., the top two parts of Table 4.2, Table 4.3 and Table 4.4). Table 4.2 and Table 4.3 show the bi-directional retrieval results for the Flickr8K and Flickr30K datasets. We implement KMM with two visual representations: VGG-based and ResNet-based. It is known that ResNet generally performs better than VGG in many tasks. As expected, KMM with ResNet-based visual representation achieves better performance than KMM with VGG-based visual representation and has an increase of 2.6 ~ 6.4%. For Flickr8K and Flickr30K, we compare our approach with 4 and 7 state-of-the-art methods, respectively. The table shows that our approach achieves better performance than the compared methods in most cases. In the task of "Text → Image", CSE achieves better results than ours in terms of the metrics R@5 and R@10. Compared with our previous work MLLM, which can be considered as a simple version of KMM that does not introduce kernel mapping, we find that KMM achieves a large improvement. The results mean that the kernel mapping may lead to better modeling for nonlinear data distributions and nonlinear relationships between modalities. Table 4.4 shows the comparison between KMM and 9 state-of-the-art methods for the MSCOCO dataset. From the table, we find that the performance of our approach is better than or close to those of the compared methods. Our approach achieves the best performance for the metrics R@5 and R@10 in the task of "Image → Text" and the metric R@10 in the task of "Text → Image", while DSvEL obtains better results than ours in the other cases.

We also evaluate our approach in the case of "1 image vs. 5 captions" and report the results in Table 4.2, Table 4.3 and Table 4.4 (i.e., the bottom two parts of the tables). The tables show that KMM is superior to MLLM and Deep CCA. Regarding Flickr30K, we find that CCL achieves better performance than our approach for the metric R@5 and R@10 in the task of "Text → Image". Comparing the case of "1 image vs. 1 caption" with "1 image vs. 5 captions", we notice that the former has a change of $-1.2 \sim +0.7$% in terms of the three recalls when KMM works with the ResNet-based visual representation. More specifically, for the task of "Text → Image", the former has a slight decline in terms of all metrics; for the task of "Image → Text", the former tends to achieve a higher recall in terms of R@1 and a lower recall in terms of R@10. We consider that this change may derive from the richness of association information between different modalities and the way

Table 4.4: Performance (percent) comparison of bi-directional retrieval on MSCOCO. The top two parts in the body of the table correspond to the case of "1 image vs. 1 caption" and the bottom two parts correspond to the case of "1 image vs. 5 captions".

Approaches	Visual Backend	Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10
HGLMM [139]	VGG	37.7	66.6	79.1	24.9	58.8	76.5
GMM+HGLMM [139]	VGG	39.4	67.9	80.9	25.1	59.8	76.6
MLLM [43]	VGG	47.6	76.3	85.2	35.9	64.3	81.6
OrderEmbedding [144]	VGG	46.7	-	88.9	37.9	-	85.9
DVSA [143]	VGG	38.4	69.9	80.5	27.4	60.2	74.8
2-Way Net [102]	VGG	55.8	75.2	-	39.7	63.3	-
DSvEL [142]	ResNet	**69.8**	91.9	96.6	**55.9**	**86.9**	94.0
CSE [145]	ResNet	56.3	84.4	92.2	45.7	81.2	90.6
Embedding Networks [103]	VGG	54.0	84.0	91.2	43.3	76.8	87.6
KMM (without sparseness)	ResNet	63.1	91.5	96.5	49.5	82.9	93.4
KMM (without smoothness)	ResNet	62.1	90.6	94.9	48.4	81.5	92.0
KMM	VGG	56.9	85.5	92.1	42.0	75.6	88.5
KMM	ResNet	63.9	**92.3**	**96.9**	50.1	83.6	**94.2**
MLLM [43] (1 ima. vs. 5 cap.)	VGG	47.3	76.1	85.6	36.8	65.5	82.3
KMM (1 ima. vs. 5 cap.)	VGG	56.3	85.1	91.8	42.9	76.1	88.7
KMM (1 ima. vs. 5 cap.)	ResNet	**63.2**	**92.0**	**97.1**	**50.8**	**84.2**	**94.6**

Table 4.5: MAP scores (percent) of bi-directional retrieval on NUS-WIDE-10K.

Approaches	Image → Text	Text → Image	Average
Deep CCA [128]	40.7	41.6	41.2
GMM+HGLMM [139]	44.0	45.3	44.7
MLLM [43]	49.7	48.1	48.9
CCL [130]	50.6	53.5	52.1
KMM (VGG)	51.7	51.6	51.7
KMM (ResNet)	**54.8**	**54.4**	**54.6**

of retrieval. Intuitively, the case of "1 image vs. 1 caption" has less association information than "1 image vs. 5 captions" due to its shorter text, hence it may result in a slightly lower recall on the whole in the image retrieval given a textual query; while for the task of "Image → Text", all 5 correct captions can be used as the candidates to match a given image query and increase the possibility of a correct one among the first r responses, especially for the metric R@1.

From Table 4.2, Table 4.3 and Table 4.4, we observe that ResNet-based representation generally leads to better performance than VGG- and AlexNet-based representations due to its better abstraction of visual semantics using the structure of more layers. Regarding the superiority of CSE, DSvEL and CCL to our approach in some cases, we consider that there are two main causes. One is the visual localization. For example, DSvEL introduces a localization mechanism to emphasize the visual concepts associated with the corresponding text. CCL uses local visual patches as well as whole images as the input of model. CSE adds the consistency constraints on the intermediate regional features. The fine grained information may help capture accurate mapping between modalities. In addition, the multi-layered association in the feature extraction via deep networks may cause the improvement. Both CCL and CSE introduce consistency constraints for images and text at different layers of deep networks, which truely reinforce the association of heterogeneous modalities.

In Table 4.5, we report the performance of bi-directional retrieval on the NUS-WIDE-10K dataset in terms of the mAP metric. Since NUS-WIDE-10K has class labels, we can compute the mAP for the retrieval task. In the experiment, we compare our approach with 4 state-of-the-art methods. As shown from the table, KMM (with ResNet-based visual representation) maintains an advantage with all 4 compared methods and KMM (with VGG-based visual representation) obtains similar results with CCL.

Performance on cross-dataset evaluation

Following RRF-Net [146] and CSE [145], we also evaluate the performance of our approach in terms of cross-dataset generalization. In this experiment, we em-

Table 4.6: Performance (percent) of bi-directional retrieval on cross-dataset in the case of "1 image vs. 1 caption".

Data Setting	Approaches	Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10
Train: Flickr30K, Test: MSCOCO	RRF-Net [146]	24.8	**53.0**	64.8	18.8	44.1	58.5
	CSE [145]	24.6	49.2	62.5	**19.1**	44.4	58.6
	KMM (ResNet)	**25.4**	52.5	**65.4**	**19.1**	**44.8**	**58.9**
Train: MSCOCO, Test: Flickr30K	RRF-Net [146]	28.8	53.8	66.4	21.3	42.7	53.7
	CSE [145]	30.6	59.3	71.0	26.0	52.1	**64.3**
	KMM (ResNet)	**32.7**	**60.1**	**71.6**	**26.6**	**52.4**	63.7

ploy the model trained on Flickr30K or MSCOCO to evaluate the test set of the other dataset. Table 4.6 reports the results of bi-directional retrieval in the case of "1 image vs. 1 caption" for the cross-dataset. The performance of the generalization is similar to and positively correlated with the performance in Table 4.2, Table 4.3 and Table 4.4. The table also shows that it is easier to transfer a model trained on a large dataset to a small one than the converse case. From the table, we observe that, on the whole, our approach achieves better performance on the cross-dataset evaluation. We consider that this may be caused by two reasons. 1) In the training process, the deep networks pre-trained on ImageNet are changeless and the feature space is uniform for different datasets. In this case, the KMM model trained in the feature space that is independent of datasets can transfer the association knowledge across datasets more stably. 2) As a model-driven approach, KMM introduces an explicit probabilistic model to describe both the data distribution and relationship distribution, which can be considered as prior information from the Bayesian viewpoint, and can generally improve generalizability.

Example illustration

Fig. 4.6 shows some examples of cross-media retrieval results in the cases of "1 image vs. 1 caption" (top two rows) and "1 image vs. 5 captions" (bottom four rows) for the MSCOCO test data. All retrieval algorithms encourage the ground truth associated with queries to be located as close to the front of the response as possible. In the first case, we find a response (in the 2nd row) that is not the ground truth associated with the query appears in front of a correct caption; in the second case, we show two examples (in the 4th and 6th rows) in which the ground truth does not appear at the 1st position in the retrieval results. We find that the retrieval results at the 1st position are truely similar with the queries. For example, in the 4th row, although the returned image at the 1st position is not the ground truth associated with the query, it consists of the same objects, such as "plane" and "runway", as the ground truth and highly matches the query.

4.6. CONCLUSIONS

In this chapter, we present a kernel-based probabilistic mixture model, called KMM, for modeling the semantic correlation between web images and text. KMM assumes that the relationship between different modalities follows multiple basic transformations, each working over a local region described by a neighborhood model in the input space. We employ kernel theory to address the nonlinearity of the data distribution and cross-modal mapping. We present a hybrid optimization algorithm based on EM and subgradient ascent to estimate the parameters of KMM and prove that the algorithm can converge to an (local) optimal solution. By combining nonlinear transformation and probabilistic models, KMM addresses the complexity of the semantic distribution over the global input space, its continuity at the local scale, and the nonlinearity in the mapping of different modalities. The experimental results demonstrate the superiority of our

Query	Retrieval results				
Three large elephants and one small elephant walking through a dusty field.					
	A door for exiting and entering the house in the kitchen.	The kitchen has a white door with a window	The back door with a window in the kitchen	A kitchen with a dishwasher double door panty and a back door	A kitchen door next to a kitchen sing and counter top
Kids playing a game of base ball while people watch. Parents watching Young boys playing baseball in the sun a young boy is at home ...					
An Aer Lingus plane touches down on an airport runway. Passenger airliner at the end of a runway waiting to take off...					
	A display in a store filled with ripe bananas. A store display that has a lot of bananas on display for sale...	A pile of oranges in crates topped with yellow bananas. There are bananas, pineapples, oranges...	A planter filled with lots of yellow and red green leafed flowers. a group of flowers sitting in a vace...	Burger with broccoli, pickle, and fork on orange plate On a plate is kept a burger and a bowl of broccoli...	A bunch of bananas sitting on top of a wooden table.A closeup of a group of bananas on a table...
	A group of people fly kites into the air on a large grassy field. Group of peo - ple outdoors flying kites...	A field full of people standing on top of a grassy field flying kites. Group of many people on a field...	A group of umbrellas together in a plaza near the Eiffel Tower. The Eiffel Tower is shown in all...	The sky is cloudy over a stop sign. A traffic sign near a high grass field near a road...	People in the water and Parachutes overhead. Many Different Sails flying over a large ...

Figure 4.6: Example cross-media retrieval results over MSCOCO test data. The top two rows correspond to the case of "1 image vs. 1 caption" and the bottom four rows correspond to the case of "1 image vs. 5 captions". Images surrounded by blue boxes and blue-colored text are ground truth. Retrieval results are arranged in decreasing order of similarity.

approach over representative state-of-the-art methods of modeling the relationships between images and text.