Pattern Recognition 72 (2017) 368–380

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Sparse Relational Topical Coding on multi-modal data

Lingyun Song*, Jun Liu, Minnan Luo, Buyue Qian, Kuan Yang

SPKLSTN Lab, Department of Computer Science and Technology, Xi'an Jiaotong University, 710049, China

ARTICLE INFO

Article history:
Received 22 March 2017
Revised 25 July 2017
Accepted 3 August 2017
Available online 4 August 2017

Keywords:
Multi-modal data
Sparse latent representation
Image annotation
Link prediction

ABSTRACT

Multi-modal data modeling lately has been an active research area in pattern recognition community. Existing studies mainly focus on modeling the content of multi-modal documents, whilst the links amongst documents are commonly ignored. However, link information has shown being of key importance in many applications, such as document navigation, classification, and clustering. In this paper, we present a non-probabilistic formulation of Relational Topic Model (RTM), i.e., Sparse Relational Multi-Modal Topical Coding (SRMMTC), to model both multi-modal documents and the corresponding link information. SRMMTC has the following three appealing properties: i) It can effectively produce sparse latent representations via directly imposing sparsity-inducing regularizers. ii) It handles the imbalance issues on multi-modal data collections by introducing regularization parameters for positive and negative links, respectively; iii) It can be solved by an efficient coordinate descent algorithm. We also explore a generalized version of SRMMTC to find pairwise interactions amongst topics. Our methods are also capable of performing link prediction for documents, as well as the prediction of annotation words for attendant images in documents. Empirical studies on a set of benchmark datasets show that our proposed models significantly outperform many state-of-the-art methods.

1. Introduction

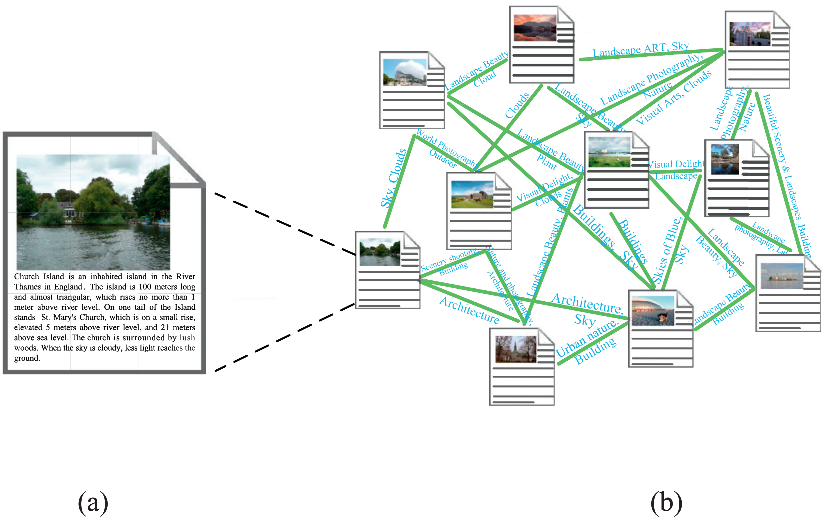
Many social media, such as Twitter, Flickr, Fotolog, and Facebook, involve multi-modal documents [1], which consist of data objects from multiple heterogeneous modalities. For example, a web page or a post from social media is a typical multi-modal document, which contains not only text, but also related images or videos. The data objects in one multi-modal document are often mutualistic in terms of delivering similar or mutually complementing high-level semantics. An example of multi-modal documents is shown in Fig. 1(a). How to formally model multi-modal documents is one of the key questions in multi-modal document understanding [2], cross-modal retrieval [3], image annotation [4–8] and classification [9–11].

Unlike some previous work in the fields of text document mining [12–14] and information retrieval [15], which usually utilizes the link information between text documents, existing work in the area of modeling multi-modal documents mainly takes the document content into account but seldom considers the link information that inherently exists between documents. Due to the interconnecting nature of the Internet, multi-modal documents are usually not independent units. In fact, they are interrelated via various explicit links (e.g., hyperlinks) or even implicit links (e.g., bibliographic citations and shared metadata). The examples of the links between multi-modal documents are shown in Fig. 1(b). The link information conveys rich semantics which are usually independent of word statistics of multi-modal documents [16]. By exploiting the link information in the modeling of multi-modal documents, we can achieve more powerful predictive models, which not only can be used to predict the links between multi-modal documents, but also can be applied to provide more accurate predictive annotation words for unknown images.

In recent years, considerable efforts have been devoted to multi-modal data modeling, which mainly can be grouped into three categories. One category of work focuses on the statistical dependency (e.g., measuring mutual information [17]) analysis of multiple modalities in a common latent space. The second category of work based on probabilistic topic models (PTMs) focuses on jointly modeling data objects with different modalities in a probabilistic manner [4,5,18]. The third category of work abstracts heterogeneous multi-modal data in a network and predicts the links between entities by utilizing the network structure and entity attributes. Although the existing approaches for modeling the multi-modal data achieve good performance, they have two limitations, which if addressed would significantly improve their performance and applicability.

* Corresponding author.
E-mail addresses: lingyun.a.song@gmail.com (L. Song), liukeen@mail.xjtu.edu.cn (J. Liu), minnluo@mail.xjtu.edu.cn (M. Luo), qianbuyue@mail.xjtu.edu.cn (B. Qian), ykxjtuedu@163.com (K. Yang).

http://dx.doi.org/10.1016/j.patcog.2017.08.005
0031-3203/© 2017 Elsevier Ltd. All rights reserved.



Church Island is an inhabited island in the River Thames in England. The island is 100 meters long and almost triangular, which rises no more than 1 meter above river level. On one tail of the Island stands St. Mary's Church, which is on a small rise, elevated 5 meters above river level, and 21 meters above sea level. The church is surrounded by lush woods. When the sky is cloudy, less light reaches the ground.

(a) (b)

Fig. 1. (a) An example of multi-modal documents. (b) Examples of links between multi-modal documents. Links exist between two multi-modal documents if they share the same metadata (e.g., groups or tags).

1) In the perspective of multi-modal document modeling, most existing studies do not account for both document content and the link information among documents, which limits their applicability and performance [2,3]. For example, some previous methods can uncover the associations between images or texts in academic articles, but they ignore the citation links amongst articles. This restricts their application in recommending related academic literature for given articles.
2) Most existing approaches cannot effectively learn sparse latent representations. Appropriate sparsity is a key factor to achieve semantically meaningful representations in document understanding tasks [19]. In addition, sparsity has also been introduced to computer vision to improve the performance of image understanding [20,21]. But only a few of existing approaches based on PTMs attempt to achieve sparsity by employing a sparse prior [22] or auxiliary variables [23]. However, due to the normalization constraints in their probabilistic formulations, the sparse representations cannot be truly achieved [19].

1.1. Overview of our model

To address the above limitations, we propose Sparse Relational Multi-Modal Topical Coding (SRMMTC). In SRMMTC, the content of multi-modal documents and the links between these documents are modeled simultaneously. A brief description of SRMMTC is presented as follows.

For multi-modal documents, SRMMTC projects data with different modalities onto a shared latent topic space and achieves nonnegative latent representations of images, words and documents. A correlation code is introduced to describe the correlations between image regions and words. SRMMTC relaxes the normalization constraints in PTMs, which brings about two advantages: i) Sparse latent representations [24,25] can be effectively learned via directly imposing appropriate regularizers. ii) The learning problem of SRMMTC can be efficiently solved by a simple coordinate descent algorithm (CDA).

For the links, SRMMTC divides them into two parts: positive links and negative links. The link between two multi-modal documents is positive when they share the same metadata (i.e., *tag* and *group*); otherwise the link is negative. For both positive and negative links, SRMMTC applies the method in [26] to impose different regularization parameters on them. This enables SRMMTC to deal with two imbalance issues: i) The imbalance between links and the words of multi-modal documents. ii) The imbalance between positive links and negative links.

Although the previous work [26] can effectively learn sparse latent representations and deal with the imbalanced issues on links, it does not take the image information of multi-modal documents into account, which limits its applications and performance.

By jointly modeling multi-modal documents and the links between them, SRMMTC can perform two kinds of predictive tasks: 1) Link prediction: Given a multi-modal document, SRMMTC can provide predictions of other multi-modal documents with which it might be linked to, using its content. 2) Image annotation: Given a multi-modal document in which descriptive words are absent, SRMMTC can provide a predictive distribution of the descriptive words for the attached image, using the links between documents and the content of image.

The results of experiments on a set of well-known benchmark datasets show that SRMMTC outperforms all the competing baseline methods.

1.2. Contributions and organization

The technical contributions of this paper can be summarized as follows:

1) We extend RTM and propose a new model to consider both the content of multi-modal documents and the links these documents, namely SRMMTC. It can be applied to solve two kinds of predictive tasks: i) Link prediction. ii) Image annotation.
2) We formulate SRMMTC as a non-probabilistic framework [19], where the normalization constraints made in PTMs are relaxed and sparse latent representations can be effectively learned through directly imposing appropriate regularizers.
3) We propose a simple CDA to efficiently solve the proposed learning problem.
4) In the tasks of link prediction and image annotation, the proposed models achieve significantly better performance compared with the competing baseline models on three well-known benchmark datasets.

The rest of this paper is organized as follows: In Section 2, we review the related work. In Section 3, we present the SRMMTC framework as a Maximum-a-Posteriori (MAP) estimate and an efficient CDA to solve it. We report the experimental results in Section 4, and finally conclude the paper in Section 5.

2. Related work

In the past decades, multi-modal data modeling has been an active research problem and a variety of models have been developed. According to the mechanisms of the model construction, these models can be typically categorized into three classes: statistical dependency modeling, probabilistic topic modeling and heterogeneous network modeling.

2.1. Statistical dependency modeling

This class of models usually maps data from different modalities into a common latent space, where the statistical dependency of different modalities is maximized. A representative work is Canonical Correlation Analysis (CCA)[27], which learns a shared subspace that maximizes the linear dependencies between different modalities. Sharma et al. [28] proposed a supervised kernelizable extension of CCA, which maps data in different modality spaces to a single (non)linear subspace. However, this type of models has two limitations: i) The latent representations of the multi-modal data obtained from the projections of these models lack apparent interpretable meanings [1]. ii) The relationships between multi-modal documents are neglected, which makes these models cannot fit real multi-modal data with natural link information.

2.2. Probabilistic topic modeling

This class of models focuses on finding conditional probabilistic relationships of different modalities data by learning their joint distribution. Representative works are mainly based on Latent Dirichlet Allocation (LDA) [22], assuming a document is a multinomial distribution over latent variables, i.e., topics, and a topic is a multinomial distribution over words. Blei and Jordan [6] extended LDA and proposed correspondence LDA (cLDA), which jointly models the distribution of images and texts and assumes that each text word directly shares a latent topic with a randomly selected image region. Putthividhy et al. [5] generalized cLDA by a regression module to correlate the topics from the different modalities. Song et al. [4] extended cLDA and proposed Sparse Multi-Modal Topical Coding (SMMTC), which can effectively find compact relations between image regions and words. The aforementioned models explicitly model the correlations between data in different modalities and discover interpretable latent representations, but they are powerless to incorporate the link relations between multi-modal documents, which would limit their applicability and performance.

Another branch of LDA-based work that focuses on modeling the link information of text documents, was proposed in recent years. For example, Chang and Blei [18] presented probabilistic Relational Topic Models (RTMs) built on LDA, in which document links are modeled as a binary random variable conditioned on the words of the documents. Zhang et al. [26] later extended RTM and proposed generalized Sparse Relational Topic Model (gSRTM) under a non-probabilistic formulation [19]. Though both the content of documents and their link information are considered, these models are originally designed for modeling mono-modal documents (i.e., texts) and cannot be applied to deal with multi-modal data. In contrast, the proposed models not only can explicitly model the correlations among different types of data, but take the link information between multi-modal documents into account, which can perform both link prediction and automatic image annotation.

In addition, another limitation of topic models is that they cannot effectively learn sparse latent representations of documents and words under normalization constraints [19]. In contrast, as a non-probabilistic formulation of Relational Topic Model, the proposed models relax the normalization constraints and can effectively learn sparse latent representations through directly using appropriate regularizers.

2.3. Heterogeneous network modeling

Another branch of models abstracts multi-modal data in a heterogeneous network. In these models, the network structure and attributes of the objects in the network are often utilized to predict the links between objects. For example, Gui et al. [29] proposed a HyperEdge-Based Embedding (HEBE) framework for a heterogeneous network consisting of multiple modality data objects. By modeling the interaction among a set of objects as a whole, HEBE preserves more contextual information to learn the latent representations of objects of different types. Sun et al. [30] proposed a meta path-based algorithm to predict the similarity links between the same type of objects in a heterogeneous information network. Shang et al. [31] proposed a framework to model heterogeneous information networks, which incorporates given meta-paths and network structures to learn latent representations of multi-modal data in the network. Chang et al. [32] proposed a deep heterogeneous Network Embedding framework, which can learn latent representations of the multi-modal data in a heterogeneous network through projecting different modalities onto a common space. These methods are robust to the data sparsity [29,33], but there is no mechanism to ensure the sparsity of the learned latent representations.

3. Sparse relational multi-modal topical coding

As a non-probabilistic topic model, SRMMTC can be formulated as a deterministic optimization problem and solved by a simple CDA.

3.1. Notation and terminology

Let $B=\{1,\cdots,M\}$ be an image vocabulary and $V=\{1,\cdots,N\}$ be a text vocabulary. By employing a bag-of-words model, an image is represented as a vector $\mathbf{r}_d=(r_{d1},r_{d2},\cdots,r_{d|I_d|})$, where $I_d\subset B$ is the index set of image words that occur and $r_{dm}(m\in I_d)$ is the number of the occurrences of image word m in the image. The corresponding caption text is represented by a vector $\mathbf{w}_d=(w_{d1},w_{d2},\cdots,w_{d|J_d|})$, where $J_d\subset V$ is the index set of text words that occur and $w_{dn}(n\in J_d)$ is the number of the occurrences of text word n.

In SRMMTC, an image and the corresponding caption text are treated as a multi-modal document. A corpus of D multi-modal documents is denoted by $C=\{(\mathbf{r}_d,\mathbf{w}_d)\}_{d=1}^{D}$, where $(\mathbf{r}_d,\mathbf{w}_d)$ represents the dth multi-modal document.

A multiset T is denoting the pairwise links between multi-modal documents in D can contain repeated elements. $\mathcal{L}=\{(d,d'):t_{d,d'}\in T\}$ denotes the set of document pairs whose links are in the training set, where $t_{d,d'}$ represents the label of the link between $(\mathbf{r}_d,\mathbf{w}_d)$ and $(\mathbf{r}_{d'},\mathbf{w}_{d'})$. Although SRMMTC can be easily extended to perform multi-type link prediction, we only consider binary links in this paper for clarity. Let $t_{d,d'}=1$ when a link exists between the dth and the d'th multi-modal document, and $t_{d,d'}=-1$ otherwise.

Let $\Psi=(\psi_{ij})\in\mathbb{R}_+{}^{K\times M}$ be an image dictionary with K topical bases, of which the kth row is denoted by $\Psi_{k\cdot}\in\mathcal{P}^M(k=1,2,\cdots,K)$, where $\mathcal{P}^M$ represents a $(M-1)$-simplex. $\Psi_{k\cdot}$ is a normalized distributional vector over all image words in B. Let $\Phi=(\phi_{ij})\in\mathbb{R}_+^{K\times M}$ be a text dictionary with K topical bases, of which the kth row is denoted by $\Phi_{k\cdot}\in\mathcal{P}^N(k=1,2,\cdots,K)$, where

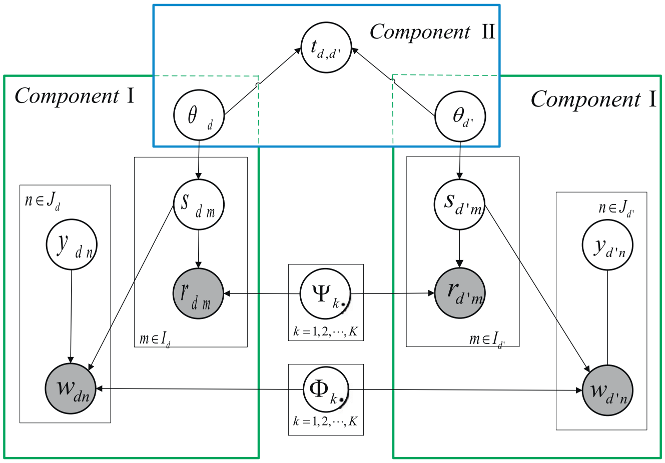


Fig. 2. A graphical model representation of SRMMTC, which only considers a pair of multi-modal documents.

$\mathcal{P}^N$ represents a $(N-1)$-simplex. $\Phi_{k.}$ is a normalized distributional vector over all text words in V.

3.2. A probabilistic generative process for SRMMTC

The graphical model representation of SRMMTC is shown in Fig. 2, which only considers one pair of multi-modal documents for the clarity of presentation. In SRMMTC, we first sample the image dictionary Ψ and the text dictionary Φ from uniform distributions on $\mathcal{P}^M$ and $\mathcal{P}^N$, respectively. Then, we model a pair of multi-modal documents C and their link T by two closely connected components.

As shown in Fig. 2, the component I (highlighted by two green boxes) focuses on modeling the content of multi-modal documents $(\mathbf{r}_d, \mathbf{w}_d)$ and $(\mathbf{r}_{d'}, \mathbf{w}_{d'})$, i.e., the images and their annotation words as well as the associations between them. For the document $(\mathbf{r}_d, \mathbf{w}_d)$, we follow [4,19] and assume that image word code $\mathbf{s}_{dm} \in \mathbb{R}_+^K (m \in I_d)$ is conditionally independent given the document code $\theta_d \in \mathbb{R}_+^K$, and assume that the observed image word count r_{dm} is independent on $\mathbf{s}_{dm}$. With the correlation code $\mathbf{y}_{dn} \in \mathbb{R}_+^{|I_d|} (n \in J_d)$, caption word code $\mathbf{z}_{dn}$ can be predicted from a linear combination of image word codes, i.e., $\mathbf{z}_{dn} = \mathbf{S}_d \mathbf{y}_{dn}$, where $\mathbf{S}_d = [\mathbf{s}_{d1}, \mathbf{s}_{d2}, \cdots, \mathbf{s}_{d|I_d|}]$[1] $\in \mathbb{R}_+^{K\times|I_d|}$. The component II (highlighted by one blue box) defines a likelihood model of the links between two multi-modal documents.

Formally, the generative procedure of SRMMTC on multi-modal documents and links between these documents can be described in the following process.

1) For the dth multi-modal document $(\mathbf{r}_d, \mathbf{w}_d)$:
   (1) Draw a multi-modal document code $\boldsymbol{\theta}_d$ from $p(\boldsymbol{\theta}_d)$.
   (2) For each observed image word $m \in I_d$:
      (a) Draw the image word code $\mathbf{s}_{dm}$ from $p(\mathbf{s}_{dm}|\boldsymbol{\theta}_d)$.
      (b) Draw the observed image word count r_{dm} from $p(r_{dm}|(\mathbf{s}_{dm})^\top \Psi_{.m})$.
   (3) For each observed caption word $n \in J_d$:
      (a) Draw the correlation code $\mathbf{y}_{dn}$ from $p(\mathbf{y}_{dn})$, and then compute caption word code by $\mathbf{z}_{dn} = \mathbf{S}_d \mathbf{y}_{dn}$.
      (b) Draw the observed caption word count w_{dn} from $p(w_{dn}|(\mathbf{z}_{dn})^\top \Phi_{.n})$.
2) For each pair of multi-modal documents $(\mathbf{r}_d, \mathbf{w}_d)$ and $(\mathbf{r}_{d'}, \mathbf{w}_{d'})$, draw a link from $p(t_{d,d'}|\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'})$.

In order to achieve the sparse codes $\boldsymbol{\theta}_d$ and $\mathbf{y}_{dn}$, we choose Laplace prior $p(\boldsymbol{\theta}_d) \propto \exp\left(-\lambda\|\boldsymbol{\theta}_d\|_1\right)$ and $p(\mathbf{y}_{dn}) \propto \exp\left(-\mu\|\mathbf{y}_{dn}\|_1\right)$. We define $p(\mathbf{s}_{dm}|\boldsymbol{\theta}_d)$ as a composite distribution

$$p\big(\mathbf{s}_{dm}|\boldsymbol{\theta}_d\big) \propto \exp\left(-\gamma\|\mathbf{s}_{dm} - \boldsymbol{\theta}_d\|_2^2 - \rho\|\mathbf{s}_{dm}\|_1\right),$$

which is super-Gaussian. The ℓ_1-norm is used to find sparse $\mathbf{s}_{dm}$ and the normal regularizer limits $\mathbf{s}_{dm}$ close to the corresponding $\boldsymbol{\theta}_d$. The parameters (λ, μ, γ, ρ) are non-negative and predefined via cross-validation.

For discrete image word count, we follow [19] to generate the observations by the Poisson distribution, i.e., $p(r_{dm}|\mathbf{s}_{dm}, \Psi) = Poiss(r_{dm}; (\mathbf{s}_{dm})^\top \Psi_{.m})$, where $Poiss(x; \nu) = \nu^x e^{-\nu}/x!$. We also generate the observation of each text word count by the Poisson distribution, i.e.,$p(w_{dn}|\mathbf{S}_d, \mathbf{y}_{dn}, \Phi) = Poiss(w_{dn}; (\mathbf{z}_{dn})^\top \Phi_{.n})$. By selecting $(\mathbf{s}_{dm})^\top \Psi_{.m}$ and $(\mathbf{z}_{dn})^\top \Phi_{.n}$ as mean parameters of the Poisson distributions, we can conveniently restrict image word codes and caption word codes to be non-negative for good interpretation [26].

For the link likelihood, we follow [18,26] and apply the sigmoid function $\sigma(x)$ to model the probability of a link,i.e.,

$$p(t_{d,d'}|\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'}) = \sigma\left(t_{d,d'}(\boldsymbol{\eta}^T(\boldsymbol{\theta}_d \circ \boldsymbol{\theta}_{d'}) + \nu)\right). \tag{1}$$

$\boldsymbol{\eta} = (\eta_1, \eta_2, \cdots, \eta_K)$ and each element η_k indicates how likely there exists a link between two multi-modal documents when they share a topic. ν represents the offset for the link probability and the notation $\circ$ denotes the Hadamard product (i.e., element-wise product).

3.3. MAP estimate

For the dth multi-modal document, we use $N_d = \{d' : (d, d') \in \mathcal{L}\}$ to denote the documents that have relationships with it in the training dataset. According to the generation procedure above, the joint distribution of SRMMTC is defined as follows:

$$p(t_{d,d'}, \mathbf{Y}_d, \mathbf{r}_d, \mathbf{w}_d, \mathbf{S}_d, \theta_d|\Phi, \Psi) = p\big(\boldsymbol{\theta}_d\big) \cdot U_d \cdot \prod_{d'\in N_d} p(t_{d,d'}|\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'}), \tag{2}$$

where $U_d = \prod_{m\in I_d} p\big(\mathbf{s}_{dm}|\boldsymbol{\theta}_d\big) p(r_{dm}|\mathbf{s}_{dm}, \Psi) \cdot \prod_{n\in J_d} p(\mathbf{y}_{dn}) p(w_{dn}|\mathbf{S}_d, \mathbf{y}_{dn}, \Phi)$ and $\mathbf{Y}_d = [\mathbf{y}_{d1}, \mathbf{y}_{d2}, \cdots, \mathbf{y}_{d|J_d|}] \in \mathbb{R}_+^{|I_d|\times|J_d|}$ represents correlation codes for text words in the dth multi-modal document.

Let $\Gamma = \{\Psi, \Phi\}$ be the dictionary for corpus C and $\Delta = \{\Theta, S, Y\}$ denote the codes for all the multi-modal documents in C. With the above joint distribution, SRMMTC can be defined as finding a MAP estimate over all the multi-modal documents. Formally, it can be formulated as solving the following optimization problem:

$$\begin{aligned} &\min_{\Delta,\Gamma,\boldsymbol{\eta},\nu} \quad \ell(S, \Psi; R) + h(S, \Phi, Y; W) + g(\Theta, \boldsymbol{\eta}, \nu; T) + \Omega(\Theta, S, Y) \\ &\text{s.t.} \quad \boldsymbol{\theta}_d \geq 0, \forall d;\ \mathbf{s}_{dm} \geq 0, \forall d, \forall m \in I_d;\ \mathbf{y}_{dn} \geq 0, \forall d, \forall n \in J_d; \\ &\qquad \Psi_{k.} \in \mathcal{P}^M;\ \Phi_{k.} \in \mathcal{P}^N, \forall k. \end{aligned} \tag{3}$$

The objective function is the negative logarithm of the posterior $p(T, Y, \Phi, \Psi, S, \Theta|R, W)$ with a constant omitted. $\Theta = \{\boldsymbol{\theta}_d\}_{d=1}^D$ represents all the multi-modal documents codes. $S = \{\mathbf{S}_d\}_{d=1}^D$ represents all the image word codes. $Y = \{\mathbf{Y}_d\}_{d=1}^D$ represents all the correlation codes. $R = \{\mathbf{r}_d\}_{d=1}^D$ and $W = \{\mathbf{w}_d\}_{d=1}^D$ denote all the images and all the caption text, respectively.

$\ell(S, \Psi; R) = -\sum_d \sum_{m\in I_d} \log Poiss(r_{dm}; (\mathbf{s}_{dm})^\top \Psi_{.m})$ is the negative log-likelihood of image word counts. $h(S, \Phi, Y; W) = -\sum_d \sum_{n\in J^d} \log Poiss\ (w_{dn};\ (\mathbf{S}_d \mathbf{y}_{dn})^\top \Phi_{.n})$ is the negative log-likelihood of text word counts. $g(\Theta, \boldsymbol{\eta}, \nu; T) = -\sum_{(d,d')\in\mathcal{L}} \log p\big(t_{d,d'}|\boldsymbol{\theta_d}, \boldsymbol{\theta_{d'}}\big)$ is the negative log-likelihood of links. The regularization term is formulated as

[1] s_{di} represents the ith image word code in the dth multi-modal document ($i = 1, 2, \cdots, |I_d|$).

$$\Omega(\Theta, S, Y) = \lambda \sum_{d=1}^{D} \|\boldsymbol{\theta}_d\|_1 + \sum_{d=1}^{D} \sum_{m \in I_d} \left(\gamma \|\mathbf{s}_{dm} - \boldsymbol{\theta}_d\|_2^2 + \rho \|\mathbf{s}_{dm}\|_1\right)$$
$$+ \mu \sum_{d=1}^{D} \sum_{n \in J_d} \|\mathbf{y}_{dn}\|_1.$$

By checking the above MAP estimate, we can find that there are two imbalanced issues: i) For each pair of multi-modal documents, there is only one link variable versus hundreds of image words and text words in them, which leads to imbalanced combination of the likelihood in the optimization probelm (3). ii) The positive links only exist between a few document pairs while most of links are negative. Therefore, in the optimization problem (3), the combination of the negative link likelihood and the positive link likelihood is imbalanced. To handle these imbalance problems, we introduce different regularization parameters for the positive and negative links respectively, which is the same as the method in [26]. The log-likelihood of links is replaced by:

$$g(\Theta, \boldsymbol{\eta}, \nu; T) = \beta_+ \sum_{(d,d') \in \mathcal{L}_+} g\big(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'}, \boldsymbol{\eta}, \nu; t_{d,d'}\big)$$
$$+ \beta_- \sum_{(d,d') \in \mathcal{L}_-} g\big(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'}, \boldsymbol{\eta}, \nu; t_{d,d'}\big),$$

where $\mathcal{L}_- = \{(d, d') \in \mathcal{L} : t_{d,d'=-1}\}$ and $\mathcal{L}_+ = \mathcal{L} \backslash \mathcal{L}_-$. The regularization parameters β_+ and β_- can be viewed as pseudo-counts of the links in terms of generative formulation. By regulating the value of β_+ and β_-, we can flexibly adjust the influence of the positive links and negative links in problem (3), which is conducive to overcome the aforementioned imbalance issues.

In the optimization problem (3), we impose non-negative constraints on $\boldsymbol{\theta}_d$, $\mathbf{s}_{dm}$ and $\mathbf{y}_{dn}$ for two reasons: i) As mean parameters of the Poisson distribution, $(\mathbf{s}_{dm})^\top \Psi_{\cdot m}$ and $(\mathbf{S}_d \mathbf{y}_{dn})^\top \Phi_{\cdot n}$ represent the reconstruction of non-negative image word counts and text word counts, respectively. ii) Non-negative constraints lead to sparser latent representations [4,19].

3.4. Optimization algorithm

In this part, we introduce the learning algorithm which solves the optimization problem (3). We set $\Gamma = \{\Psi, \Phi\}$ and use $\Delta = \{\Theta, S, Y\}$ to represent the inferred codes of the multi-modal documents in C. This optimization problem is bi-convex, i.e., convex over Δ given Γ and the parameters $\boldsymbol{\eta}$ and ν; and convex over Γ, $\boldsymbol{\eta}$ and ν given Δ. This biconvex problem can be solved by using a CDA, which is typically used in sparse coding methods [4,19,26].

We summarize the CDA for SRMMTC in Algorithm 1, which alternatively solves four subproblems. To be more specific, the first subproblem is introduced in Section 3.4.1, which learns multi-modal document codes, image word codes and correlation codes. The second and the third subproblems are described in Section 3.4.2, which learn the image dictionary and text dictionary, respectively. In Section 3.4.3, we present the fourth subproblem which aims to learn the link likelihood model.

Algorithm 1 CDA for SRMMTC.

```
Input: multi-modal documents corpus $C$.
Output: $\Theta, S, Y, \eta, \nu$.
Initialize $\Theta, S, \Phi, \Psi, \eta, \nu$ and $Y$,
while not converge do
    $(\Theta, S, Y) = Hierarchical\_Sparse\_Coding(\Phi, \Psi, \boldsymbol{\eta}, \nu)$;
    $\Psi = Image\_Dictionary\_Learning(S)$;
    $\Phi = Text\_Dictionary\_Learning(S, Y)$;
    $(\boldsymbol{\eta}, \nu) = LinkLikelihood\_Learning(\Theta)$;
end while.
```

3.4.1. Hierarchical sparse coding for multi-modal documents

In this step, image word codes S, multi-modal document codes Θ and correlation codes Y are alternatively computed with given Γ, $\boldsymbol{\eta}$ and ν. Since multi-modal documents are independent of each other, we can perform this step for each document separately. For notation simplicity, we set $\beta = \beta_- = \beta_+$. The optimization problem is

$$\min_{\theta_d, \mathbf{S}_d, \mathbf{Y}_d} \sum_{m \in I_d} \ell(\mathbf{s}_{dm}, \Psi) + \sum_{n \in J_d} h(\mathbf{S}_d, \mathbf{y}_{dn}, \Phi)$$
$$+ \beta \sum_{d' \in N_d} g\big(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'}, \boldsymbol{\eta}, \nu; t_{d,d'}\big) + \Omega\big(\boldsymbol{\theta}_d, \mathbf{S}_d, \mathbf{Y}_d\big)$$
$$\text{s.t.} \quad \boldsymbol{\theta}_d \geq 0; \ \mathbf{s}_{dm} \geq 0, \forall m \in I_d; \ \mathbf{y}_{dn} \geq 0, \forall n \in J_d, \tag{4}$$

which can be solved by applying a similar coordinate descent method as in [19,26].

According to the sigmoid link function, the negative log-likelihood of links is

$$g(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'}, \boldsymbol{\eta}, \nu; t_{d,d'}) = log\big(1 + exp(-t_{d,d'}(\boldsymbol{\eta}^T(\boldsymbol{\theta}_d \circ \boldsymbol{\theta}_{d'}) + \nu))\big). \tag{5}$$

The regularization term with respect to the dth multi-modal document is

$$\Omega\big(\boldsymbol{\theta}_d, \mathbf{S}_d, \mathbf{Y}_d\big) = \lambda \|\boldsymbol{\theta}_d\|_1 + \sum_{m \in I_d} \left(\gamma \|\mathbf{s}_{dm} - \boldsymbol{\theta}_d\|_2^2 + \rho \|\mathbf{s}_{dm}\|_1\right)$$
$$+ \mu \sum_{n \in J_d} \|\mathbf{y}_{dn}\|_1.$$

To be more specific, we alternatively update θ_d, $\mathbf{S}_d$ and $\mathbf{Y}_d$ by solving the following three subproblems.

(1) Optimization over $\boldsymbol{\theta_d}$: when $\mathbf{S}_d$ is fixed, $\boldsymbol{\theta_d}$ can be obtained by solving the convex problem

$$\min_{\boldsymbol{\theta}_d} \lambda \|\boldsymbol{\theta}_d\|_1 + \gamma \sum_{m \in I_d} \|\mathbf{s}_{dm} - \boldsymbol{\theta}_d\|_2^2 + \beta \sum_{d' \in N_d} g(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'}, \boldsymbol{\eta}, \nu; t_{d,d'})$$
$$\text{s.t.} \quad \boldsymbol{\theta}_d \geq 0. \tag{6}$$

The objective function with respect to each element of $\boldsymbol{\theta}_d$ is convex when other elements are fixed, and we alternatively solve each θ_{dk} by employing a coordinate descent method which has been used in many sparse coding methods [19,34,35]. Due to the nonlinearity of the sigmoid likelihood, this subproblem does not have a closed-form solution. It can be solved by applying a projected gradient descent algorithm. To be more specific, we first carry out a gradient descent step with line search, and then project the solution onto the convex feasible domain. Let $\mathcal{M}$ be the objective function of problem (6). Therefore, each θ_{dk} is solved by

$$[\theta_{dk}]^{new} = P([\theta_{dk}]^{old} - t\nabla_{\theta_{dk}}\mathcal{M}),$$

where t is a step size; P is a projection operator which takes care of the non-negative constraints; and the gradient of $\mathcal{M}$ w.r.t. $\boldsymbol{\theta}_{dk}$ is

$$\nabla_{\theta_{dk}}\mathcal{M} = \lambda + 2\gamma\left(\theta_{dk} \cdot |I_d| - \sum_{n \in |I_d|} s_{nk}\right) + \beta \sum_{d' \in N_d} \frac{\partial g(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'}, \boldsymbol{\eta}, \nu; t_{d,d'})}{\partial \theta_{dk}}. \tag{7}$$

The last item in Eq. (7) is the derivative of the sigmoid link function in Eq. (5) w.r.t. to θ_{dk}, which is computed by

$$\frac{\partial g(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'}, \boldsymbol{\eta}, \nu; t_{d,d'})}{\partial \theta_{dk}} = \frac{-t_{d,d'} exp(z_{d,d'})}{1 + exp(z_{d,d'})} \cdot \eta_k \theta_{d'k},$$

where $z_{d,d'} = -t_{d,d'}(\boldsymbol{\eta}^T(\boldsymbol{\theta}_d \circ \boldsymbol{\theta}_{d'}) + \nu)$.

When updating each θ_{dk}, the complexity of the gradient computation is $O(K + a)$, where $a = \max_{d \in \{1,2,\cdots,D\}} |I_d|$. As there are K

elements in each document code, the complexity of computing each multi-modal document code is $O(K^2+Ka)$.

(2) Optimization over $\mathbf{s}_{dm}$: we alternatively compute each $\mathbf{s}_{dm}$ when $\boldsymbol{\theta}_d$, $\mathbf{y}_{dn}$ and other elements of $\mathbf{S}_d$ are fixed by solving the following optimization problem.

$$\begin{aligned}\min_{\mathbf{s}_{dm}}\ & \ell(\mathbf{s}_{dm},\Psi)+\sum_{n\in J_d} h(\mathbf{S}_d,\mathbf{y}_{dn},\Phi)+\gamma\|\mathbf{s}_{dm}-\boldsymbol{\theta}_d\|_2^2+\rho\|\mathbf{s}_{dm}\|_1\\ \text{s.t.}\ \ & \mathbf{s}_{dm}\geq 0, \forall m\in I_d. \end{aligned} \tag{8}$$

Let $\mathscr{L}$ be the objective function, which is convex over each element of $\mathbf{s}_{dm}$ when other elements are fixed. We solve each element of $\mathbf{s}_{dm}$ alternatively using a projected gradient descent algorithm. The gradient of $\mathscr{L}$ w.r.t. $\mathbf{s}_{dmk}$ is

$$\begin{aligned}\nabla_{\mathbf{s}_{dmk}}\mathscr{L}=&\left(1-\frac{r_{dm}}{(\mathbf{s}_{dm})^{\top}\Psi_{\cdot m}}\right)\Psi_{km}\\ &+\sum_{n\in J_d}\left(1-\frac{w_{dn}}{(\mathbf{z}_{dn})^{\top}\Phi_{\cdot n}}\right)y_{dnm}\Phi_{kn}+2\gamma(s_{dmk}-\theta_{dk})+\rho,\end{aligned}$$

where $\mathbf{z}_{dn}=\mathbf{S}_d\mathbf{y}_{dn}$ and y_{dnm} denotes the mth element of $\mathbf{y}_{dn}$.

When updating each image word, the complexity of the gradient computation is $O(Kb)$, in which $b=\max_{d\in\{1,2,\cdots,D\}}|J_d|$. As there are $|I_d|$ image words in the dth multi-modal document, the complexity of computing all the image word codes is $O(Kab)$.

(3) Optimization over $\mathbf{y}_{dn}$: given all the image word codes $\mathbf{s}_{dm}$, we can compute each correlation code $\mathbf{y}_{dn}$ by optimizing problem

$$\begin{aligned}\min_{\mathbf{y}_{dn}}\ & h(\mathbf{S}_d,\mathbf{y}_{dn},\Phi)+\mu\|\mathbf{y}_{dn}\|_1\\ \text{s.t.}\ \ & \mathbf{y}_n^d\geq 0,\end{aligned} \tag{9}$$

which is the log-Poisson loss of caption word counts with a ℓ_1-norm regularizer.

As correlation codes y_{dn} is not coupled, we can optimize each correlation code separately. Let $\mathcal{H}=h(S_d,\mathbf{y}_{dn},\Phi)+\mu\sum_{m\in I_d}y_{dnm}$ be the objective function. Then we compute each y_{dnm} alternatively when other elements of y_{dn} are fixed. The solution is $y_{dnm}=\max(0,\upsilon_m)$, where $\upsilon_m=\arg\min_{y_{dnm}}\mathcal{H}$ with $y_{dni}(i\in I_d\backslash\{m\})$ fixed at current solutions [19]. Setting the gradient

$$\begin{aligned}\nabla_{y_{dnm}}\mathcal{H}\ =&\ (1-\frac{w_{dn}}{\mathbf{y}_{dn}^{\top}\mathbf{S}_d^{\top}\Phi_{\cdot n}})\mathbf{s}_{dm}^{\top}\Phi_{\cdot n}+\mu\\ =&\left(1-\frac{w_{dn}}{\sum_{i\in I_d\backslash\{m\}}\sum_{j=1}^{K}y_{dni}s_{dij}\phi_{jn}+y_{dnm}\sum_{j=1}^{K}s_{dmj}\phi_{jn}}\right)\mathbf{s}_{dm}^{\top}\Phi_{\cdot n}+\mu=0,\end{aligned}$$

then, we have $y_{dnm}=\max(0,\upsilon_m)$ with

$$\upsilon_m=\left(\frac{w_{dn}}{1+\frac{\mu}{\mathbf{s}_{dm}^{\top}\Phi_{\cdot n}}}-u_1\right)/u_2, \tag{10}$$

where $u_1=\sum_{i\in I_d\backslash\{m\}}\sum_{j=1}^{K}y_{dni}s_{dij}\phi_{jn}$, $u_2=\sum_{j=1}^{K}s_{dmj}\phi_{jn}$, s_{dij} represents the jth element of $\mathbf{s}_{di}$ and s_{dmj} represents the jth element of $\mathbf{s}_{dm}$. The complexity of computing each correlation code y_{dn} is $O(Ka)$. As the dth multi-modal document has $|J_d|$ correlation codes, the complexity of computing all the correlation codes is $O(Kab)$.

3.4.2. Dictionaries learning for images and text

In this step, we compute the image dictionary Ψ and the text dictionary Φ with given codes $\Delta=\{\Theta,S,Y\}$ for all the multi-modal documents. Ψ and Φ can be solved by minimizing the log-Poisson loss $\ell(S,\Psi;R)$ and $h(S,\Phi,Y;W)$, respectively. $\ell(S,\Psi;R)$ and $h(S,\Phi,Y;W)$ is convex with respect to the corresponding dictionary and each dictionary is constrained on a probabilistic simplex. Therefore, we naturally use a projected gradient descent algorithm to update each dictionary and then project each row of the dictionary onto an ℓ_1-simplex [19]. The complexity of updating Ψ and Φ are $O(KaD)$ and $O(KabD)$, respectively.

3.4.3. Link likelihood learning

After inferring the multi-modal document codes Θ, we can compute the link likelihood parameters $\boldsymbol{\eta}$ and ν. In this part, we only need to consider the link part $\mathcal{G}=\beta\sum_{(d,d')\in\mathcal{L}}g\big(\boldsymbol{\theta}_d,\boldsymbol{\theta}_{d'},\boldsymbol{\eta},\nu;t_{d,d'}\big)$ and the optimization probelm is:

$$\min_{\boldsymbol{\eta},\nu}\beta\sum_{(d,d')\in\mathcal{L}}\log\left(1+exp(-t_{d,d'}(\boldsymbol{\eta}^T(\boldsymbol{\theta}_d\circ\boldsymbol{\theta}_{d'})+\nu))\right), \tag{11}$$

which is convex for $\boldsymbol{\eta}$ and v. We find the solutions for parameters η and v by employing a gradient descend method. The gradient of $\mathcal{G}$ w.r.t. θ_{η_k} is

$$\nabla_{\eta_k}\mathcal{G}=\beta\sum_{(d,d')\in\mathcal{L}}\frac{-t_{d,d'}\theta_{dk}\theta_{d'k}exp(z_{d,d'})}{1+exp(z_{d,d'})},$$

and its complexity is $O(K|\mathcal{L}|)$. The gradient of $\mathcal{G}$ w.r.t. $\boldsymbol{\theta}_v$ is

$$\nabla_{\nu}\mathcal{G}=\beta\sum_{(d,d')\in\mathcal{L}}\frac{-t_{d,d'}exp(z_{d,d'})}{1+exp(z_{d,d'})}$$

and its complexity is $O(|\mathcal{L}|)$.

3.4.4. Complexity

As shown in Algorithm 1, the time complexity of SRMMTC is determined by four parts: Hierarchical_Sparse_Coding, Image_Dictionary_Learning, Text_Dictionary_learning and LinkLikelihood_Learning.

In Hierarchical_Sparse_Coding, the image word code $\mathbf{s}_m^d$, the multi-modal document code $\boldsymbol{\theta}^d$ and the correlation code $\mathbf{y}_n^d$ are alternatively computed. Therefore, the complexity of Hierarchical_Sparse_Coding over D multi-modal documents is $O(D\cdot(K^2+Ka+Kab+Kab))=O(D\cdot(K^2+Kab))$.

The image dictionary Ψ and the text dictionary Φ are updated by a projected gradient descent algorithm. The complexity of updating Ψ and Φ are $O(KaD)$ and $O(KabD)$, respectively.

The Linklikelihood_Learning alternatively solves parameters η and v by employing a gradient descent method. The complexity of updating η and v are $O(K|\mathcal{L}|)$ and $O(|\mathcal{L}|)$.

In summary, the complexity of SRMMTC is $O(t\cdot(KabD+K^2D+KabD+KaD+K|\mathcal{L}|+|\mathcal{L}|))=O(K^2tD+KabtD+K|\mathcal{L}|t)$, where t denotes the iteration number of the *while* loop in Algorithm 1.

3.5. A generalized sparse relational multi-modal topic model

In SRMMTC, though $\boldsymbol{\eta}^T(\boldsymbol{\theta}_d\circ\boldsymbol{\theta}_{d'})+\nu$ describes the strength of the connections between two multi-modal documents, it only captures the same topic interactions. That is, only the topic shared by two documents can make a contribution to their link likelihood. However, different latent topics also have correlations with each other [36], which have influence on document links and if captured would lead to better link prediction results [26]. It is more likely to find a link between two multi-modal documents which are on closely related topics. Therefore, we follow [26] and generalize SRMMTC by using a full weight matrix $L^{K\times K}$, aiming to capture all topic interactions. The link likelihood is redefined as:

$$p(t_{d,d'}|\boldsymbol{\theta}_d,\boldsymbol{\theta}_{d'})=\sigma\left(t_{d,d'}(\boldsymbol{\theta}_d^{\top}L\boldsymbol{\theta}_{d'}+\nu)\right). \tag{12}$$

where L_{ij} denotes the connection strength of two documents when they have topic i and topic j respectively. In this generalized SRMMTC (gSRMMTC) model, we also adopt the sigmoid likelihood function for the links. The optimization problem of the link likelihood is similar to that of SRMMTC, and can be solved by using a similar CDA. Only a few changes are needed to make in the gradient descent steps of learning document codes and link likelihood.

4. Experiment

In this part, we first introduce the experimental datasets in Section 4.1. Then, we apply the proposed models (i.e., SRMMTC and gSRMMTC) to the task of link prediction and show the experimental results in Section 4.2. To get insights into the properties of the proposed models, we also present sensitivity analyses on parameters. Finally, we compare the proposed models with the baseline models on the task of image annotation in Section 4.3.

4.1. Datasets

We adopt three well-known multi-modal benchmark datasets used in [37], which contain large numbers of paired images and texts as well as social network metadata. The images and their metadata are gleaned from Flickr, such as image tags, image groups and the locations where the images are created. In this paper, we view a pair of image and annotation text as a multi-modal document and establish the links between those documents by using the metadata. According to McAuley and Leskovec [37], we know that documents sharing even a single group or tag are much more likely to have semantic relationships. Therefore, to build strong semantic relationships between multi-modal documents, we build a positive link between two multi-modal documents when the attached images share the same groups and tags. The three datasets used in our experiments are introduced below:

1) The ImageCLEF dataset [38] consists of 4,546 multi-modal documents, and on average, each image in the documents is annotated with approximately 12 words from a text vocabulary of 145 terms in total. There exist 21,192 tags and 10,575 groups, and each image is annotated with approximately ten tags and five groups averagely. Among the multi-modal documents there are 49,404 positive links. So on average each multi-modal document has approximately eleven links to other documents and the ratio of positive links is approximate to 0.2%.
2) The MIR dataset [39] contains 14,460 multi-modal documents with 51,040 tags, 21,894 groups and a text vocabulary of 159 terms. There are 391,731 positive links among those documents, so only 0.2% of document pairs have positive links. On average, each image in the documents is associated with approximately three annotation words from the vocabulary, ten tags and five groups.
3) The PASCAL dataset [40] is another multi-modal document dataset consisting of 10,189 documents and 82,854 positive links among those documents. The ratio of document pairs that have positive links in this dataset is only approximately 0.07%. A total of 27,250 tags and 6,951 groups exist in the entire dataset. On average, each image is annotated with approximately seven tags and two groups. The text vocabulary of this dataset consists of 50 words.

All datasets are divided into 5 folds, with the goal of performing a 5-fold cross-validation. We preprocess the images in those multi-modal documents as in [4,5]. To be more specific, we first resize the image to the size of 224×224 and segment each training image into some 20×20 patches. Then, we extract 128-dim SIFT descriptors and 36-dim robust color descriptors [41] from the 20×20 gray-scale patches. Finally, we quantize the combined 164-dim features into 256 clusters via using *k*-means. Each center of clusters is treated as a word in the image vocabulary.

4.2. Link prediction

4.2.1. Performance evaluation

Given a new multi-modal document, we follow [18,26] and predict its links to the training documents. This prediction can be done in three steps by applying the proposed models. First, we infer the latent representation of each testing document via a hierarchical sparse coding step. Second, we compute the probability of existing a link between two documents by the logistic link likelihood function. Finally, according to a link's probability, we can make a binary decision. That is, if the probability is larger than 0.5, there exists a link; otherwise, no link exists.

To the best of our knowledge, there are few attempts in the literature to predict the links of multi-modal documents. Therefore, we only select the following two representative models as the comparative models.

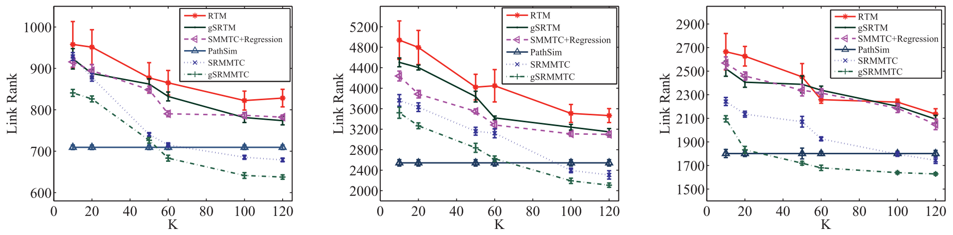
1) RTM [18]: a heuristic model extends LDA for link prediction and has been extensively selected as a baseline [26,42]. In this model, each document is first modeled as in LDA. Then the links between the documents are modeled as binary variables, of which the distribution is defined by a logistic link function. Note that RTM only focuses on the text content of documents and no image information is used when it performs link prediction.
2) SMMTC + Regression: a two-step model first trains a SMMTC [4] to discover the sparse latent representations of multi-modal documents. Then, it learns a logistic regression model on training links to predict the links of test documents. Note that the link information does not have effects on the latent representations of multi-modal documents.
3) gSRTM [26]: a non-probabilistic topic model utilizes both text content of multi-modal documents and link information when performing link prediction. gSRTM can handle the imbalance issues on links and learn sparse latent representations of documents and words. It overlooks the image information inside multi-modal documents when predicting the links between documents.
4) PathSim [30]: a meta-path based link prediction algorithm aims to find the similarity links between the same type of objects in heterogeneous networks. It has been extensively used to predict the links among peer objects and achieved superior performance [31].

For probabilistic RTM, we follow the setup in [18] and only use observed links as training data. For sparse topic models, including SRMMTC, gSRMMTC, gSRTM and the de-coupled approach of SMMTC + Regression, we randomly draw 0.5% of the unobserved links as negative examples, which can partly redress the serious imbalance issues and decrease the computational cost of inference [26,43]. Note that different sampling ratios (e.g., 0.1% and 0.2%) do not have much influence on the link prediction results of the proposed models, because they can effectively handle the imbalance issues by tuning the regularization parameters β_+ and β_-.

For PathSim, we construct the heterogeneous network consisting of three types of objects D, T and G. *D* denotes multi-modal documents. *T* denotes the tags associated with each multi-modal document. *G* denotes groups that multi-modal document belongs to. Links exist between either *D* and *T*, or *D* and *G*. We follow [30] to evaluate the performance of PathSim under the meta path *DTDGDTD*, which implies two multi-modal documents could be related to each other without the same tags shared, as long as their tags are used by many other multi-modal documents within the same groups.

We use *link rank* [26,43] as the performance measure, which is the same as that in [18]. For a test document, its *link rank* is defined as the average rank of the observed links between it and training documents. The general *link rank* is an average of the *link rank* over all test documents. A lower value of *link rank* implies a better prediction performance.

For better performance evaluation, we use 5-fold cross-validation. Parameters of all the models are tuned to their best



(a) ImageCLEF dataset (b) MIR dataset (c) PASCAL dataset

Fig. 3. Link ranks of different models with different values of K on different datasets.

settings for link prediction. For RTM, we tune the Dirichlet hyper-parameter α. For the proposed models, we set $\rho = \gamma$ and perform cross-validation to select λ, γ and μ.

Fig. 3 shows the 5-fold cross-validation results of link prediction on all datasets with respect to the variation of topic numbers. As seen from the figure, the proposed models perform better than all other competitors on all datasets. ImageCLEF is paradigmatic, showing a nearly 18% improvement in *link rank* for gSRMMTC over RTM ,13.4% improvement over SMMTC + Regression and 14.3% improvement over gSRTM. SRMMTC performs nearly as well on ImageCLEF with an approximately 13.1% improvement over RTM, 8.2% improvement over SMMTC + Regression and 9.1% improvement over gSRTM. The proposed models also achieve better performance than PathSim when $K > 100$.

The superior performance of the proposed models over RTM can be attributed to three reasons:

1) Besides the text content of multi-modal documents, the proposed models also take the attendant image information into account when predicting its links. Whereas RTM performs the prediction based on the text content only.
2) The proposed models can flexibly address the imbalance issues by introducing appropriate regularization parameters for both positive links and negative links. However, RTM is powerless in dealing with the imbalance issues [26,43].
3) The proposed models are effective in learning sparse code representations of documents, which is consistent with the intrinsic sparsity of topic meanings for a document [19,23,26]. In contrast, RTM can only have an indirect impact on the sparsity of posterior representations through using a Dirichlet prior, which has been demonstrated inefficient [19].

The superior performance of the proposed models over SMMTC + Regression shows the benefits of the involvement of link information for link prediction. By incorporating link information during the hierarchical sparse coding step, the proposed models can discover more discriminative document representations for link prediction. In contrast, SMMTC + Regression learns the latent representations of multi-modal documents without any link information, basing on the document content only.

The superior performance of the proposed models over gSRTM shows the benefits of the consideration of multiple modalities (i.e., text and images) involved, which is important to learn representations of multi-modal documents [44].

The superior performance of the proposed models over PathSim can be attributed to two reasons: i) Besides the link information among multi-modal documents, the proposed models exploit the content of multi-modal documents and explore the correlations among the multiple modalities inside these documents, which are important to model multi-modal data [44]. In contrast, PathSim mainly applies meta-paths to guide the prediction of the links among multi-modal documents. ii) The proposed models can effectively learn sparse latent representations of multi-modal documents, which is important to improve the performance of link prediction [26].

The superior performance of gSRMMTC over the diagonal SRMMTC demonstrates the significance of capturing all pairwise topic interactions. By using a full weight matrix, gSRMMTC can capture valuable topic relationships, which are useful in improving the performance of link prediction [26].

Table 1 illustrates suggested multi-modal documents that have links for the testing document shown in the first row of the table, using gSRMMTC, RTM and SMMTC+Regression as predictive models. These suggestions were computed from a model trained on the ImageCLEF dataset. We can see that gSRMMTC outperforms SMMTC+Regression and RTM owing to more correct links identified. For the testing document, gSRMMTC finds three related documents, only one related document is found by both RTM and SMMTC+Regression. The truly related documents are highlighted in Table 1 with boldfaced text words.

To verify the robustness of the proposed models when data become sparse, we randomly sample different ratios (10%, 20%, 30%, 50%, 70%) of the three datasets to repeat link prediction experiments. The experimental results of different models are reported in Table 2 for the ImageCLEF dataset, Table 3 for the MIR dataset, and Table 4 for the PASCAL dataset. The link sparsity ratios are defined as the proportions of document pairs that have positive links and vary with different sampling ratios. The ratios are reported in the first two rows of the tables. Based on the vertical comparison from Tables 2–4, we observe that the proposed models achieve the best performance in all cases and are relatively more robust to data sparsity. This is because the proposed models can flexibly handle the imbalance issues on links by introducing regularization parameters for positive and negative links, respectively.

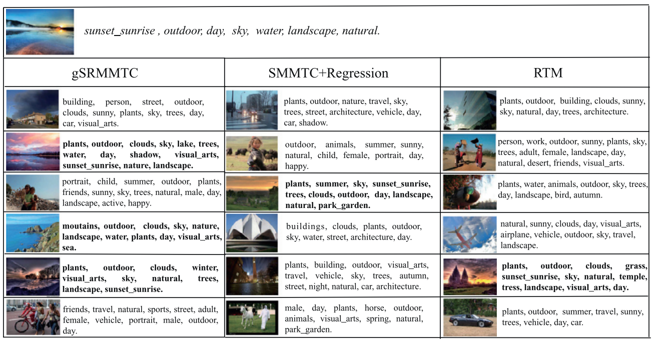
4.2.2. Sensitivity analyses on parameters

We perform sensitivity analyses on their parameters to demonstrate their advantages of learning sparse codes and dealing with the imbalance issues. We also analyze how these parameters influence the link prediction performance.

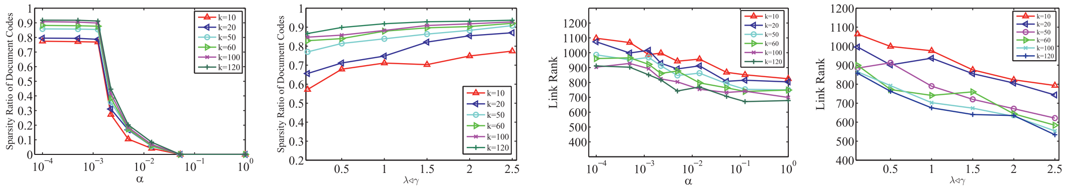
Document code sparsity By tuning the parameters (ρ, λ, γ), the sparsity level of multi-modal document codes learned by the proposed models can be flexibly controlled, which is an advantage of the proposed models. Recalling the optimization problem (4), we can see that the proposed models bias towards learning sparse document codes when setting λ to a relatively large value. Being similar to the work of [4,26], we fix $\rho = \gamma$ and tune the ratio λ/γ. For RTM, the Dirichlet hyper-parameters α control the sparsity level of document representations. We follow [26] and use common symmetric Dirichlet prior for the topic mixing proportions in RTM.

Fig. 4(a) and (b) show the variation of the sparsity ratio of document codes discovered by RTM and SRMMTC on the ImageCLEF dataset when tuning their parameters. Sparsity ratio is defined as the average ratio of zero elements in codes. As RTM does not produce zero code elements [19,26], we truncate a minuscule value

Table 1
Top 6 link predictions made by gSRMMTC, SMMTC + Regression and RTM for one multi-modal document (text words are italicized) from ImageCLEF. Boldfaced text words indicate actual related multi-modal documents.



sunset_sunrise , outdoor, day, sky, water, landscape, natural.		
gSRMMTC	SMMTC+Regression	RTM
building, person, street, outdoor, clouds, sunny, plants, sky, trees, day, car, visual_arts.	plants, outdoor, nature, travel, sky, trees, street, architecture, vehicle, day, car, shadow.	plants, outdoor, building, clouds, sunny, sky, natural, day, trees, architecture.
plants, outdoor, clouds, sky, lake, trees, water, day, shadow, visual_arts, sunset_sunrise, nature, landscape.	outdoor, animals, summer, sunny, natural, child, female, portrait, day, happy.	person, work, outdoor, sunny, plants, sky, trees, adult, female, landscape, day, natural, desert, friends, visual_arts.
portrait, child, summer, outdoor, plants, friends, sunny, sky, trees, natural, male, day, landscape, active, happy.	**plants, summer, sky, sunset_sunrise, trees, clouds, outdoor, day, landscape, natural, park_garden.**	plants, water, animals, outdoor, sky, trees, day, landscape, bird, autumn.
moutains, outdoor, clouds, sky, nature, landscape, water, plants, day, visual_arts, sea.	buildings, clouds, plants, outdoor, sky, water, street, architecture, day.	natural, sunny, clouds, day, visual_arts, airplane, vehicle, outdoor, sky, travel, landscape.
plants, outdoor, clouds, winter, visual_arts, sky, natural, trees, landscape, sunset_sunrise.	plants, building, outdoor, visual_arts, travel, vehicle, sky, trees, autumn, street, night, natural, car, architecture.	**plants, outdoor, clouds, grass, sunset_sunrise, sky, natural, temple, tress, landscape, visual_arts, day.**
friends, travel, natural, sports, street, adult, female, vehicle, portrait, male, outdoor, day.	male, day, plants, horse, outdoor, animals, visual_arts, spring, natural, park_garden.	plants, outdoor, summer, travel, sunny, trees, vehicle, day, car.



(a) Sparsity of document codes by RTM (b) Sparsity of document codes by SRMMTC (c) Link Rank using RTM (d) Link Rank using SRMMTC

Fig. 4. sparsity ratios (a) and link ranks (c) for RTM with different number of topical bases on the ImageCLEF dataset; sparsity ratios (b) and link ranks (d) for SRMMTC with different number of topical bases on the ImageCLEF dataset.

Table 2
Link ranks of different models on sampled ImageCLEF data.

Sampling ratio Link Sparsity ratio	10% 0.034%	20% 0.061%	30% 0.079%	50% 0.117%	70% 0.157%
RTM	1097.8	1033.2	922.5	893.2	882.8
gSRTM	1028.2	968.5	931.5	871.7	865.5
SMMTC + Regression	992.6	946.8	906.4	867.2	837.3
PathSim	882.2	827.8	792.7	768.5	752.8
SRMMTC	810.2	774.8	745.2	724.4	706.2
gSRMMTC	772.1	758.6	719.3	692.7	675.4

Table 3
Link ranks of different models on sampled MIR data.

Sampling ratio Link Sparsity ratio	10% 0.029%	20% 0.051%	30% 0.069%	50% 0.096%	70% 0.127%
RTM	4560.8	4421.3	4135.6	3985.4	3728.1
gSRTM	4035.4	3841.7	3686.1	3521.2	3451.7
SMMTC + Regression	3994.6	3808.3	3653.6	3475.2	3465.9
PathSim	3307.9	3140.8	2902.7	2799.5	2730.3
SRMMTC	2889.8	2759.4	2707.6	2609.6	2562.2
gSRMMTC	2604.2	2550.8	2433.4	2357.7	2311.3

Table 4
Link ranks of different models on sampled PASCAL data.

Sampling ratio Link Sparsity ratio	10% 0.012%	20% 0.019%	30% 0.026%	50% 0.039%	70% 0.052%
RTM	2936.8	2808.7	2691.6	2595.3	2347.5
gSRTM	2820.4	2680.7	2465.3	2365.6	2294.9
SMMTC + Regression	2671.4	2498.7	2328.4	2224.6	2198.5
PathSim	2260.8	2136.6	2023.1	1983.5	1923.3
SRMMTC	2151.9	2076.2	1972.1	1898.6	1850.3
gSRMMTC	2109.2	1942.6	1869.5	1825.7	1789.3

ϵ (e.g., $\epsilon < 0.001$) to be zero. Correspondingly, we show the variations of *link rank* of RTM and SRMMTC in Fig. 4(c) and(d) when tuning the parameters, respectively. For SRMMTC, we fix γ to a constant when tuning the ratio λ/γ and for RTM we tune the Dirichlet parameter α.

As seen from Fig. 4(a) and (c), RTM learns sparse document codes by using a small α, but its *link rank* is very high. That is, RTM does not perform well in link prediction at this sparsity. Although a relatively large α boosts the performance of link prediction, it leads to a dramatic drop in the sparsity ratio. Fig. 4(c) shows that RTM performs the best when the sparsity ratio is zero. This demonstrates that a small α in the Dirichlet prior of RTM cannot effectively yield sparse posterior representations.

In contrast, SRMMTC achieves sparser document codes by using the sparsity-inducing ℓ_1 regularizer on document codes. In Fig. 4(b), we can see that for SRMMTC the sparsity ratio of inferred multi-modal document codes is high and stable in a wide range ($1 < \lambda/\gamma \leq 2.5$). When the number of topical bases is relatively small, the sparsity ratio can be gradually improved through increasing λ. In addition to the sparsity ratio, the increase of λ also causes a gradual descent in the *link rank* which can be seen in Fig. 4(d). Its best *link rank* is acquired at a relatively high sparsity ratio when λ/γ is around 2.5. This suggests that sparser multi-modal document representations bring about better link prediction performance.

The effects of parameter β As discussed in Section 3.3, there are two imbalance issues of modeling multi-modal documents and their links, which are the imbalance between modeling words(i.e., image words and text words) and links, and the imbalance between positive links and negative links. The proposed models can effectively deal with these issues by regulating the values of parameters β_+ and β_-, which are introduced into positive links and negative links, respectively. For the first issue, we set β_+ and β_- to appropriate values, which can improve the influence of links in the optimization problem (3). The second issue can be handled by tuning the ratio β_+/β_-, because a relatively large β_+/β_- can decrease the influence of negative links while increasing the influence of

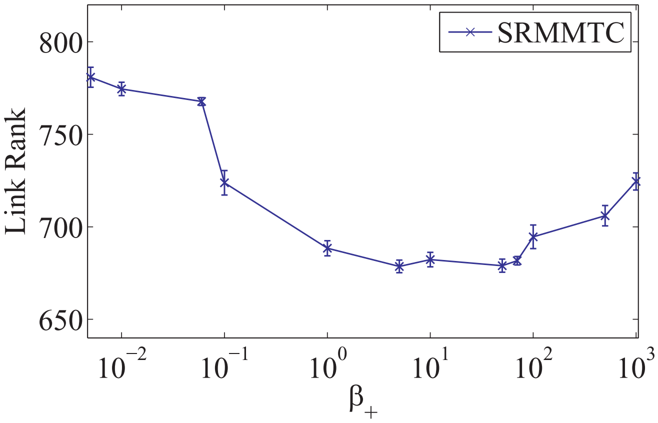


Fig. 5. Link rank of SRMMTC using different values of β_+ on the ImageCLEF dataset.

positive links. Since the sub-sampling strategy used in [26] is useful to improve the results of link prediction, we sub-sample 0.5% negative links as the training data.

Fig. 5 shows link ranks of SRMMTC with different β_+ values when the number of topical bases K is fixed at 120. We follow [26] and set $\beta_+ = 10\beta_-$ to balance the positive links and negative links. As seen from the figure, SRMMTC achieves better performance when $5 < \beta_+ < 80$. This is consistent with the formulation of optimization problem (4). When β_+ is too small (e.g., $\beta_+ < 0.01$), the influence of the link information on the objective function of optimization problem (4) is small, meanwhile image words and text words dominate the objective function. On the contrary, the link information dominates the whole objective function when β_+ is larger than 70. With moderate β_+ and β_- values, SRMMTC can utilize both the content of multi-modal documents and their link information rather than only one of them, which leads to better performance.

4.3. Automatic image annotation

The second predictive task is automatic image annotation, where we predict the absent words of a multi-modal document using its accompanied links and image content. Following the baseline models [4–6], we adopt *caption perplexity* as the performance measure. *Perplexity* is computed by

$$perplexity = \exp\left\{-\sum_{d=1}^{D_{test}}\sum_{n\in J_d}\log p(n|\mathbf{r}_d)/\sum_{d=1}^{D_{test}}|J_d|\right\}, \quad (13)$$

where D_{test} represents the number of test multi-modal documents and $p(n|\mathbf{r}_d)$ represents the conditional probability of the text word n given the image $\mathbf{r}_d$. Since the *perplexity* is equal to the inverse of the geometric mean likelihood, lower perplexity indicates better prediction performance.

Following Song et al. [4], we can obtain $p(n|\mathbf{r}_d)$ by using the cosine similarity between the image code and the text word code. To be more specific, we first infer the image word codes $\mathbf{s}_{dm}(m \in J_d)$ through hierarchical sparse coding. Then, we compute the image code by $\mathbf{s}_d = \sum_{m\in I_d} r_{dm}\mathbf{s}_{dm}$, where r_{dm} is the code weight of the image word m. Finally, we compute the word code $\bar{\mathbf{z}}_n$ of n by $\bar{\mathbf{z}}_n = (1/|C_n|)\sum_{d\in C_n}\mathbf{z}_{dn}$, where C_n is the index set of multi-modal documents in which the word n occurs.

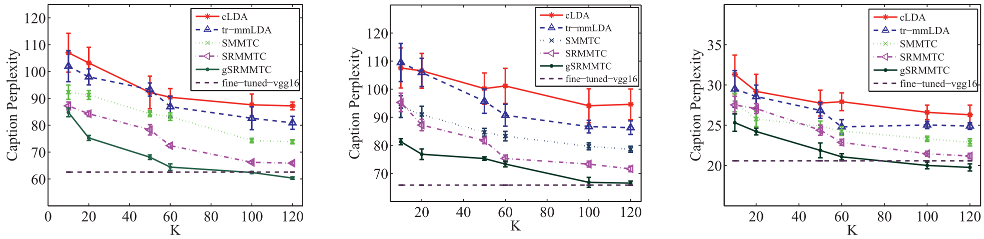
To demonstrate the advantage of the proposed models in automatic image annotation, we compare it with four baselines on several benchmark datasets. Note that All baseline models perform image annotation without using any link information. The details are presented as follows:

1) VGG-16 [45]: a very popular deep learning model, which has been applied to many image recognition tasks and achieved superior performance. However, learning a deep model usually requires a large number of labeled image samples. The limited amount of labeled training data in our datasets will hinder its performance. To overcome this difficulty, we fine-tune it as [46] does. To be more specific, we freeze partial convolutional layers of a VGG-16 network pre-trained on ImageNet, fine-tune the remaining convolutional layers and fully connected layers, then train the classifier layer using our experimental data.
2) cLDA [6]: a popular latent variable model extends LDA for image annotation and has been extensively used as a baseline [4,5]. It is effective in modeling the conditional distribution of the annotation words given an image.
3) tr-mmLDA [5]: a well-known extension of cLDA, in which a regression module is introduced to regulate the correlations between image regions and annotation words.
4) SMMTC [4]: a non-probabilistic topic model that extends STC [19] for image annotation. SMMTC associates each text word with only a few semantically related image regions rather than all the image regions indistinguishably.

The 5-fold cross-validation results of image annotation are presented in Fig. 6, which shows the variations of caption perplexity obtained by each model with different values of K. As seen from the figure, when the pre-defined parameter $K > 100$ we can draw the following two conclusions: i) The perplexities obtained by the proposed models is the lowest compared with baselines on the ImageCLEF dataset and the PASCAL dataset. It suggests that the proposed models can achieve superior predictive performance on these datasets. Although there is no significant difference between the performance of the proposed model gSRMMTC and the fine-tuned VGG-16 on the MIR dataset, gSRMMTC performs better than the fine-tuned VGG-16 on the ImageCLEF dataset and the PASCAL dataset. Another advantage of the proposed models is reflected in its low requirement for the size of training data. They do not need be pre-trained but can achieve good performance with limited labeled training data (e.g., thousands of labeled images). In contrast, to achieve the comparative performance, VGG-16 requires massive amounts of labeled training data (e.g., millions of labeled images), which limits its applicability. ii) The perplexities of two PTMs (i.e.,cLDA and tr-mmLDA) are constantly higher than the non-probabilistic topic models (i.e., SMMTC, SRMMTC and gSRMMTC). That is, the non-probabilistic topic models perform better than the two PTMs. The superior performance of the proposed models is attributed to two reasons:

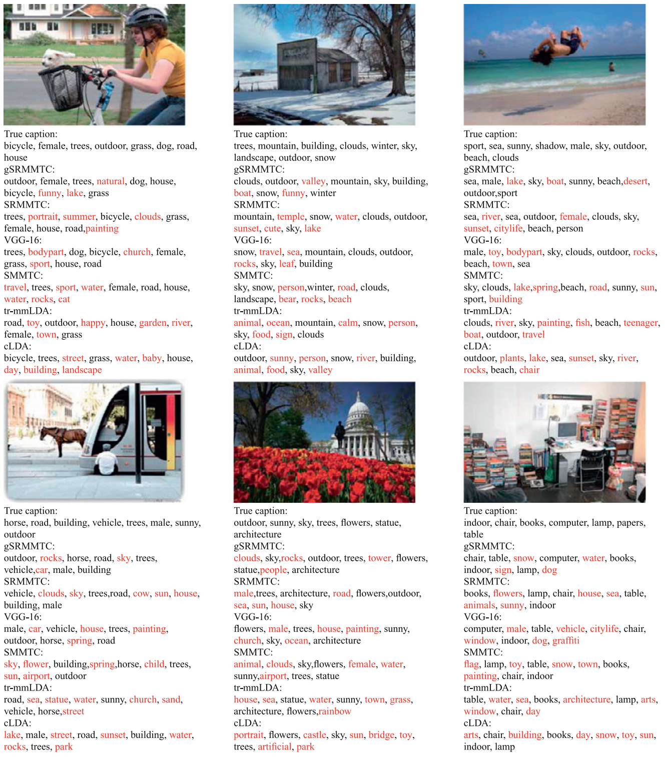
1) The proposed models can effectively learn sparse latent representations. The sparsity can lead to compact and high-fidelity representations and has been demonstrated an important property in many computer vision tasks [4,20]. This can partly explains why the proposed models outperform the PTMs and VGG-16, which are inefficient in learning sparse latent representations.
2) The proposed models take into account the link information that exists in multi-modal data collections, which is very useful to improve the performance of image annotation [37]. The importance of link information can be seen from the comparisons between the proposed models and SMMTC. Although SMMTC has the same advantage of learning sparse latent representations, it does not use any link information in predicting annotation words as the proposed models do.

We show in Fig. 7 six example images from the test set of ImageCLEF and their annotations predicted by each model when $K = 120$. The incorrect predictions are highlighted by using red font. As seen from the figure, the proposed models achieve the best performance and label most of the example images correctly.



(a) ImageCLEF dataset (b) MIR dataset (c) PASCAL dataset

Fig. 6. Caption perplexities with different values of K on different datasets.


True caption:
bicycle, female, trees, outdoor, grass, dog, road, house
gSRMMTC:
outdoor, female, trees, natural, dog, house, bicycle, funny, lake, grass
SRMMTC:
trees, portrait, summer, bicycle, clouds, grass, female, house, road,painting
VGG-16:
trees, bodypart, dog, bicycle, church, female, grass, sport, house, road
SMMTC:
travel, trees, sport, water, female, road, house, water, rocks, cat
tr-mmLDA:
road, toy, outdoor, happy, house, garden, river, female, town, grass
cLDA:
bicycle, trees, street, grass, water, baby, house, day, building, landscape

True caption:
trees, mountain, building, clouds, winter, sky, landscape, outdoor, snow
gSRMMTC:
clouds, outdoor, valley, mountain, sky, building, boat, snow, funny, winter
SRMMTC:
mountain, temple, snow, water, clouds, outdoor, sunset, cute, sky, lake
VGG-16:
snow, travel, sea, mountain, clouds, outdoor, rocks, sky, leaf, building
SMMTC:
sky, snow, person,winter, road, clouds, landscape, bear, rocks, beach
tr-mmLDA:
animal, ocean, mountain, calm, snow, person, sky, food, sign, clouds
cLDA:
outdoor, sunny, person, snow, river, building, animal, food, sky, valley

True caption:
sport, sea, sunny, shadow, male, sky, outdoor, beach, clouds
gSRMMTC:
sea, male, lake, sky, boat, sunny, beach,desert, outdoor,sport
SRMMTC:
sea, river, sea, outdoor, female, clouds, sky, sunset, citylife, beach, person
VGG-16:
male, toy, bodypart, sky, clouds, outdoor, rocks, beach, town, sea
SMMTC:
sky, clouds, lake,spring,beach, road, sunny, sun, sport, building
tr-mmLDA:
clouds, river, sky, painting, fish, beach, teenager, boat, outdoor, travel
cLDA:
outdoor, plants, lake, sea, sunset, sky, river, rocks, beach, chair

True caption:
horse, road, building, vehicle, trees, male, sunny, outdoor
gSRMMTC:
outdoor, rocks, horse, road, sky, trees, vehicle,car, male, building
SRMMTC:
vehicle, clouds, sky, trees,road, cow, sun, house, building, male
VGG-16:
male, car, vehicle, house, trees, painting, outdoor, horse, spring, road
SMMTC:
sky, flower, building,spring,horse, child, trees, sun, airport, outdoor
tr-mmLDA:
road, sea, statue, water, sunny, church, sand, vehicle, horse,street
cLDA:
lake, male, street, road, sunset, building, water, rocks, trees, park

True caption:
outdoor, sunny, sky, trees, flowers, statue, architecture
gSRMMTC:
clouds, sky,rocks, outdoor, trees, tower, flowers, statue,people, architecture
SRMMTC:
male,trees, architecture, road, flowers,outdoor, sea, sun, house, sky
VGG-16:
flowers, male, trees, house, painting, sunny, church, sky, ocean, architecture
SMMTC:
animal, clouds, sky,flowers, female, water, sunny,airport, trees, statue
tr-mmLDA:
house, sea, statue, water, sunny, town, grass, architecture, flowers,rainbow
cLDA:
portrait, flowers, castle, sky, sun, bridge, toy, trees, artificial, park

True caption:
indoor, chair, books, computer, lamp, papers, table
gSRMMTC:
chair, table, snow, computer, water, books, indoor, sign, lamp, dog
SRMMTC:
books, flowers, lamp, chair, house, sea, table, animals, sunny, indoor
VGG-16:
computer, male, table, vehicle, citylife, chair, window, indoor, dog, graffiti
SMMTC:
flag, lamp, toy, table, snow, town, books, painting, chair, indoor
tr-mmLDA:
table, water, sea, books, architecture, lamp, arts, window, chair, day
cLDA:
arts, chair, building, books, day, snow, toy, sun, indoor, lamp

Fig. 7. Example images from the test set of ImageCLEF and their annotations predicted by different models.

5. Conclusions and future work

In this paper, we have proposed SRMMTC and gSRMMTC for modeling multi-modal documents. The proposed models account for not only the correlations among different modalities, but also the link information between multi-modal documents. By relaxing the restrictive normalization constraints in PTMs, the proposed models obtain two properties: i) They can effectively learn sparse latent representations. ii) They can be efficiently solved by a CDA. In addition, the proposed models can deal with the imbalanced issues existing in collections of multi-modal documents. Experimental results demonstrate that our proposed models outperform all the competing baseline models on the tasks of link prediction and image annotation.

In the future, we will build new multi-modal datasets containing multiple types of links and discuss the robustness of the prediction models on links of different types.

Acknowledgements

The research was supported in part by National Key Research and Development Program of China under Grant No. 2016YFB1000903; National Science Foundation of China under Grant Nos. 61672419, 61672418, 61532004, 61502377, 61532015; The MOE Research Program for Online Education under Grant No. 2016YB166; The Fundamental Research Funds for the Central Universities; The Ministry of Education Innovation Research Team No. IRT_17R86; The Project of China Knowledge Centre for Engineering Science and Technology. We would like to give our appreciation to Yihe Chen, an undergraduate from University of Toronto, who give us many suggestions in the writing style of formal English language.

References

[1] Y. Wang, F. Wu, J. Song, X. Li, Y. Zhuang, Multi-modal mutual topic reinforce modeling for cross-media retrieval, in: Proceedings of the 22nd ACM International Conference on Multimedia, 2014, pp. 307–316.
[2] Y. Yi, Z. Yue-Ting, W. Fei, P. Yun-He, Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval, IEEE Trans. Multimedia 10 (3) (2008) 437–446.
[3] S. Wang, Y. Wu, Q. Huang, Improving cross-modal correlation learning with hyperlinks, in: IEEE International Conference on Multimedia and Expo, 2015, pp. 1–6.
[4] L. Song, M. Luo, J. Liu, L. Zhang, B. Qian, M.H. Li, Q. Zheng, Sparse multi-modal topical coding for image annotation, Neurocomputing 214 (2016) 162–174.
[5] D. Putthividhy, H.T. Attias, S.S. Nagarajan, Topic regression multi-modal latent dirichlet allocation for image annotation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3408–3415.
[6] D.M. Blei, M.I. Jordan, Modeling annotated data, in: Proceedings of the 26th international ACM SIGIR conference on Research and development in informaion retrieval, 2003, pp. 127–134.
[7] Z. Li, D. Gong, X. Li, D. Tao, Aging face recognition: a hierarchical learning model based on local patterns selection, IEEE Trans. Image Process. 25 (5) (2016) 2146–2154.
[8] F. Wang, J. Wang, C. Zhang, J. Kwok, Face recognition using spectral features, Pattern Recognit. 40 (10) (2007) 2786–2797.
[9] C. Gong, D. Tao, S.J. Maybank, W. Liu, G. Kang, J. Yang, Multi-modal curriculum learning for semi-supervised image classification, IEEE Trans. Image Process. 25 (7) (2016) 3249–3260.
[10] J. Yu, D. Tao, Y. Rui, J. Cheng, Pairwise constraints based multiview features fusion for scene classification, Pattern Recognit. 46 (2) (2013) 483–496.
[11] Y. Zheng, J. Fan, J. Zhang, X. Gao, Hierarchical learning of multi-task sparse metrics for large-scale image classification, Pattern Recognit. 67 (2017) 97–109.
[12] X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, J. Han, Cluscite: Effective citation recommendation by information network-based clustering, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 821–830.
[13] H. Gui, Y. Sun, J. Han, G. Brova, Modeling topic diffusion in multi-relational bibliographic information networks, in: Proceedings of the 23th ACM International Conference on Conference on Information and Knowledge Management, 2014, pp. 649–658.
[14] R.M. Nallapati, A. Ahmed, E.P. Xing, W.W. Cohen, Joint latent topic models for text and citations, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 542–550.
[15] J. Tang, J. Zhang, J.X. Yu, Z. Yang, K. Cai, R. Ma, L. Zhang, Z. Su, Topic distributions over links on web, in: IEEE International Conference on Data Mining, 2009, pp. 1010–1015.
[16] M. Henzinger, Hyperlink analysis on the world wide web, in: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia, 2005, pp. 1–3.
[17] A. Papoulis, S.U. Pillai, Probability, Random Variables, and Stochastic Processes, Tata McGraw-Hill Education, 2002.
[18] J. Chang, D.M. Blei, Relational topic models for document networks, in: International Conference on Artificial Intelligence and Statistics, volume 9, 2009, pp. 81–88.
[19] J. Zhu, E. Xing, Sparse topical coding, in: Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence, 2011, pp. 831–838.

[20] W. Liu, D. Tao, J. Cheng, Y. Tang, Multiview hessian discriminative sparse coding for image annotation, Comput. Vis. Image Understanding 118 (2014) 50–60.

[21] L. Zhang, Q. Zhang, L. Zhang, D. Tao, X. Huang, B. Du, Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding, Pattern Recognit. 48 (10) (2015) 3102–3112.

[22] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[23] C. Wang, D.M. Blei, Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process, in: Advances in Neural Information Processing Systems, 2009, pp. 1982–1989.

[24] M. Luo, F. Sun, H. Liu, Hierarchical structured sparse representation for t–s fuzzy systems identification, IEEE Trans. Fuzzy Syst. 21 (6) (2013) 1032–1043.

[25] M. Luo, F. Sun, H. Liu, Joint block structure sparse representation for multi--input–multi-output (MIMO) t–s fuzzy system identification, IEEE Trans. Fuzzy Syst. 22 (6) (2014) 1387–1400.

[26] A. Zhang, J. Zhu, B. Zhang, Sparse relational topic models for document networks, in: Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2013, pp. 670–685.

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

[28] A. Sharma, A. Kumar, H. Daume, D.W. Jacobs, Generalized multiview analysis: a discriminative latent space, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2160–2167.

[29] H. Gui, J. Liu, F. Tao, M. Jiang, B. Norick, J. Han, Large-scale embedding learning in heterogeneous event data, in: IEEE International Conference on Data Mining, 2016, pp. 907–912.

[30] Y. Sun, J. Han, X. Yan, P.S. Yu, T. Wu, Pathsim: meta path-based top-k similarity search in heterogeneous information networks, Proceedings of the VLDB Endowment 4 (11) (2011) 992–1003.

[31] J. Shang, M. Qu, J. Liu, L.M. Kaplan, J. Han, J. Peng, Meta-path guided embedding for similarity search in large-scale heterogeneous information networks, arXiv preprint arXiv:1610.09769 (2016).

[32] S. Chang, W. Han, J. Tang, G.-J. Qi, C.C. Aggarwal, T.S. Huang, Heterogeneous network embedding via deep architectures, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 119–128.

[33] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, Q. Mei, Line: large-scale information network embedding, in: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 1067–1077.

[34] S. Bengio, F. Pereira, Y. Singer, D. Strelow, Group sparse coding, in: Advances in neural information processing systems, 2009, pp. 82–89.

[35] H. Lee, R. Raina, A. Teichman, A.Y. Ng, Exponential family sparse coding with application to self-taught learning., in: IJCAI, vol. 9, 2009, pp. 1113–1119.

[36] D. Blei, J. Lafferty, Correlated topic models, Adv. Neural Inf. Process. Syst. 18 (2006) 147.

[37] J. McAuley, J. Leskovec, Image labeling on a network: using social-network metadata for image classification, in: European Conference on Computer Vision, 2012, pp. 828–841.

[38] S. Nowak, M.J. Huiskes, New strategies for image annotation: overview of the photo annotation task at imageCLEF, CLEF LABs and Workshops, 2010.

[39] M.J. Huiskes, M.S. Lew, The MIR flickr retrieval evaluation, in: ACM Sigmm International Conference on Multimedia Information Retrieval, 2008, pp. 39–43.

[40] M. Everingham, L.V. Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

[41] J.V.D. Weijer, C. Schmid, Coloring local feature extraction, in: Proceedings of European Conference on Computer Vision, vol. 2, 2006, pp. 334–348.

[42] H. Wang, X. Shi, D.-Y. Yeung, Relational deep learning: a deep latent variable model for link prediction, in: Association for the Advancement of Artificial Intelligence, 2017, pp. 2688–2694.

[43] N. Chen, J. Zhu, F. Xia, B. Zhang, Generalized relational topic models with data augmentation, in: International Joint Conference on Artificial Intelligence, 2013, pp. 1273–1279.

[44] T. Lu, Y. Jin, F. Su, P. Shivakumara, C.L. Tan, Content-oriented multimedia document understanding through cross-media correlation, Multimed. Tools Appl. 74 (18) (2015) 8105–8135.

[45] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.

[46] M. Long, Y. Cao, J. Wang, M.I. Jordan, Learning transferable features with deep adaptation networks, in: International Conference on Machine Learning, 2015, pp. 97–105.

Lingyun Song received the M.S. degree in Software Engineering from Xi'an Jiaotong University, China, in 2014. Currently, he is working toward the Ph.D. degree in the Department of Computer Science at Xi'an Jiaotong University, China. His research interests include machine learning, pattern recognition, image understanding.

Jun Liu received the B.S., M.S., and Ph.D. degrees from Xi'an Jiaotong University, China, in 1995, 1998, and 2004, respectively, all in computer science. He is currently a professor in the Department of Computer Science, Xi'an Jiaotong University.

Minnan Luo received her Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2014. Currently, she is an Assistant Professor in the School of Electronic and Information Engineering at Xi'an Jiaotong University.

Buyue Qian is currently an Associate Professor at Xi'an Jiaotong University. He received his Ph.D. in 2013 from Computer Science Department, University of California at Davis. Before that, he received Master of Science (2009) from Columbia University, and B.S. in Information Engineering (2007) from Xi'an Jiaotong University.

Kuan Yang is working toward the B.S. degree in the Department of Computer Science at Xi'an Jiaotong University, China. His research interests include data mining, pattern recognition, image understanding and searching.