# 融合多层语义的跨模态检索

冯耀功 蔡国永

(桂林电子科技大学计算机与信息安全学院 广西 桂林 541004)

**摘 要** 如何挖掘出不同模态数据之间的潜在语义关联是跨模态检索算法的核心问题。已有研究表明,将表示学习和关联学习融合的模式比较适用于跨模态检索的任务,但目前基于这一模式的模型的不同模态数据的抽象层次之间只包含着1-1的对应关联关系。由于异构多模态数据的抽象粒度并不完全相同,对此它们之间的关联关系很可能不只存在于指定的抽象层上。因此,提出了一种融合多层语义的跨模态检索模型,它利用深度玻尔兹曼机的双向结构特点,实现了将文本模态数据的不同抽象层次同时关联到图像模态数据的多个抽象层上,从而更充分地挖掘不同模态数据抽象层之间 *N*-*M* 的内在关联。基于3个公开数据集的实验结果表明,该模型优于之前类似的跨模态检索模型,具有更高的检索精确度。

**关键词** 深度学习,跨模态,检索,多层语义,融合

**中图法分类号** TP183 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.03.034

## Cross-modal Retrieval Fusing Multilayer Semantics

FENG Yao-gong CAI Guo-yong

(School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China)

**Abstract** How to explore the inherent relations of different modalities is the core problem of cross-modal retrieval. The previous works demonstrate that the models which incorporate representation learning and correlation learning into a single process are more suitable for cross-modal retrieval task, but these models only contain the 1-1 correspondence correlations between different modalities. However, different modalities are more likely to have different granularities of semantics abstraction, and the correlations between different modalities are more likely to occur in different layers of semantic at the same time. This paper proposed a cross-modal retrieval model fusing multilayer semantic. The model benefits from the architecture of deep boltzmann machine which is an undirected graph model and implements that each semantic layer of text modality is associated with multiple different semantic layers of image modality at last, and explores the inherent *N*-*M* relations of different modalities more sufficiently. The results of experiments on three real and public datasets demonstrate that this model is obviously superior to the state-of-art models, and has higher accuracy of retrieval.

**Keywords** Deep learning, Cross-modal, Retrieval, Multilayer semantics, Fusion

## 1 引言

在检索信息时,常出现试图使用一种模态的数据查询其他模态数据的情况,例如使用文本描述查找图片,或者使用图片查找描述文本,因此产生了跨模态检索的需求。在跨模态检索领域,多模态数据的建模过程通常包含表示学习和关联学习两部分。表示学习通常是将单模态数据进行抽象化表示,又被称为单模态表示学习。关联学习通常是指基于不同模态之间数据的对应关系,通过典型关联分析(Canonical Correlation Analysis,CCA)或者共享表示层等类似的技术,将不同模态的数据映射到一个公共的表示空间中,从而进一步利用某种距离函数度量出不同模态数据之间的相似性[1]。已有研究发现,将单模态数据的表示学习和挖掘多模态数据之间内在关联的关联学习相融合的多模态数据建模方式更加适合跨模态检索的任务[2-6]。

Feng等提出的Corr-AE模型[2]和栈式Corr-RBMs等模型[3]就是基于这一模式的。在Corr-AE模型中,Feng等将包含了对应抽象层次相似性约束的关联学习与自编码器(AutoEncoder,AE)相结合;在栈式Corr-RBMs模型中,他们将类似的关联学习与受限玻尔兹曼机(Restricted Boltzmann Machine,RBM)相结合。这两个模型分别构建了单层次和多层次的对应关联关系。而Wang等[4]将这种包含相似性约束的关联学习融入到了栈式自编码器中,构建了MSAE模型。Cai等[5]则将这种包含相似性约束的关联学习同时融入到多

到稿日期:2018-02-07 返修日期:2018-05-16 本文受国家自然科学基金(61763007),广西自然科学基金(2017JJD160017)资助。

**冯耀功**(1992—),男,硕士,主要研究方向为跨模态检索,E-mail:fengyaogong@gmail.com;**蔡国永**(1971—),男,博士,教授,主要研究方向为社交媒体数据挖掘,E-mail:ccgycai@gmail.com(通信作者)。

种神经网络中，利用各种神经网络的优势进行互补，构建了DCN模型。Peng等[6]融合了多种神经网络，提出了CMDN模型，将基于共享表示层的关联学习和表示学习统一在模型学习的第一阶段，然后在第二阶段将两者的结果向量进行拼接；需要指出的是，在第二阶段的训练过程中，CMDN模型需要利用带标签的数据对网络进行微调，从而进行有监督的训练。

虽然将表示学习和关联学习相融合的模式更适合跨模态检索的任务，但是目前这类模型依然存在局限性。在这类模型中，只包含了人为指定的不同模态数据之间的单个或者多个层次的1-1的对应关联关系，但是这不足以挖掘出异构多模态数据之间的潜在语义一致性；而且，从数据本身的角度来看，不同模态的数据本身的抽象粒度不尽相同，并且文本数据相较于图像数据通常有着更加明确的语义指向性[3]。因此很可能出现如下情况：文本数据的某个抽象层次中所包含的语义内容不仅与图像数据对应抽象层的语义内容相关，还与图像数据更高抽象层的局部语义内容和更低抽象层的整体语义内容相关。文献[7]也指出为这类模型设计一种更为复杂的多模态数据之间的关联关系，从而挖掘更细粒度的异构模态数据的潜在语义关联是未来工作所要努力的方向。

为了打破上述局限，本文使用具有双向特性的深度玻尔兹曼机[8](Deep Boltzmann Machine，DBM)来构建异构多模态数据之间更为复杂的关联关系，将DBM与包含相似性约束的多层次的跨模态关联学习融合为一个整体，提出了一种无监督的融合多层语义的跨模态检索模型(Cross-modal Retrieval that Fusing Multilayer Semantics，CRFMS)。利用DBM在数据建模过程中双向传递的特性[8-9]，使文本模态的每个抽象层次和图像模态数据的多个抽象层次同时构建起多个关联关系，希望能更好地挖掘出不同模态数据之间的潜在关联语义。

本文第1节介绍了基于几个经典的将表示学习和关联学习相融合的跨模态检索模型；第2节详细描述了本文所提出的CRFMS模型；第3节通过3个真实公开数据集上的实验描述了CRFMS模型较已有模型存在的优势；最后总结全文。

## 2 CRFMS模型

### 2.1 模型结构

CRFMS模型的框架如图1所示，其由两类DBM以及每个抽象层次之间的关联关系组成。

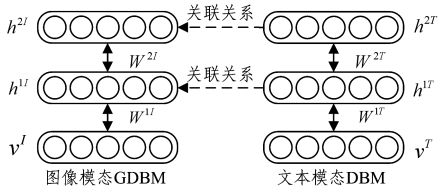



图1 CRFMS框架

Fig. 1 Structure of CRFMS

图1中左半部分是适合图像数据建模的GDBM(Gaussian DBM)[10]，由底层的高斯玻尔兹曼机[11-12](Gaussian RBM，GRBM)和顶层的标准二值RBM(standard binary RBM)组成；右半部分是适合为文本建模的DBM，由底层的神经主题模型[13](Replicated Softmax RBM，RSRBM)和顶层的标准二值RBM组成。

因为DBM是一个标准的无向图模型[8]，包含了自下而上的传递和自上而下的反馈两个通道，所以DBM存在双向传递的特性；每个抽象层次的分布同时依赖其上一层和下一层的分布，即每个抽象层次不仅可以接收到来自底层抽象信息的传递，而且还能接收到来自更高层次的抽象信息的反馈，从而生成更好的依赖于上下层特征的表示。为了能够让文本模态数据的某个抽象层次与图像模态数据的多个抽象层次同时构建多个关联关系，CRFMS在$h^1$层和$h^2$层建立了单向的对应关联关系，即由文本抽象层指向图像抽象层的关联关系。每次通过这种对应关联关系发生的各自层次的状态变化，都可以同时影响到这个图像抽象层的更高抽象层次和更低抽象层次的变化。通过这种方式，使得文本模态数据的某个抽象层次与图像模态数据的多个抽象层次同时构建了多个关联关系，从而更加充分地挖掘出文本模态和图像模态数据之间的潜在语义关联，提升了跨模态检索效果。

### 2.2 训练算法

CRFMS模型的训练算法是在DBM训练算法[14]的基础上设计的。该算法将包含相似性约束的关联学习融入到了DBM的训练过程中，使得CRFMS模型在融合了数据多个抽象层语义的同时，在不同模态数据之间建立起了更为复杂的关联关系。

CRFMS模型的训练过程如图2所示，该模型从整体上可以分为逐层预训练阶段(上半部分)和单向融合阶段(下半部分)。

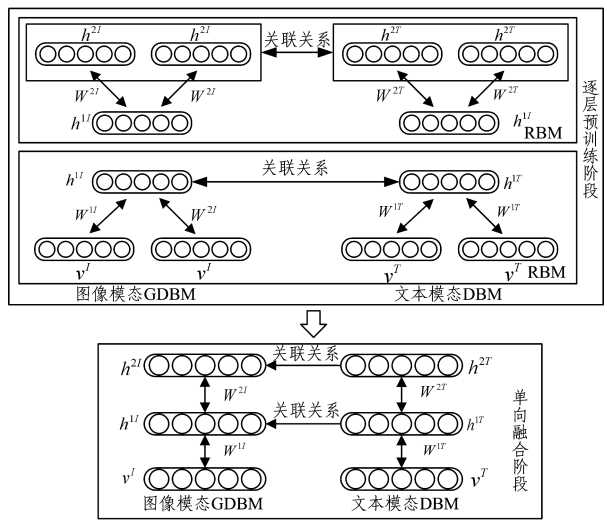



图2 CRFMS模型的训练过程

Fig. 2 Training process of CRFMS

(1)逐层预训练阶段

预训练阶段采用对比散度(Contrastive Divergence，CD)算法[15]逐层训练CRFMS。但是，此过程中的CD算法过程与一般的CD算法过程是有区别的[14]。如图2中CRFMS预训练部分的第一层所示，两个RBM的可见层单元数量增加

一倍，通过这样的变化来弥补缺少的自上而下的反馈，从而确定隐层单元的状态。因此，依据训练过程的改变，CRFMS 模型第一层的可视单元和隐藏单元的条件概率分布如式(1)和式(2)所示：

$$p(h_j^1=1|v)=\sigma(\sum_i W_{ij}^1 v_i+\sum_i W_{ij}^1 v_i) \tag{1}$$

$$p(v_i=1|h^1)=\sigma(\sum_i W_{ij}^1 h_j) \tag{2}$$

基于上述条件概率分布，第一层的目标函数如式(3)所示：

$$\min L^1=L_D^1+\alpha L_I^1+\beta L_T^1 \tag{3}$$

其中：

$$L_D^1=\sum_{j=1}\| h_j^{1I}-h_j^{1T}\|_2^2 \tag{4}$$

$$L_I^1=-\sum_{i=1}\log p(v_i^I) \tag{5}$$

$$L_T^1=-\sum_{i=1}\log p(v_i^T) \tag{6}$$

其中，$L_D$ 表示两种模态数据之间的相似性约束，即第 $i$ 组图像向量和文本向量在统一表示空间中的欧氏距离之和，本文称之为不同模态数据之间的关联误差；$L_I$ 和 $L_T$ 表示图像数据和文本数据的负对数似然值[15]；$\alpha$ 和 $\beta$ 是控制图像和文本数据在整个目标函数中所占的权重超参数。

CRFMS 模型第一层的训练过程如下：首先使用 CD 算法来对 CRFMS 第一层的 GRBM 和 RSRBM 进行训练，更新 GRBM 和 RSRBM 的输入层偏置项、第一层隐层偏置项和它们之间的权重值；然后使用关联误差再次对 CRFMS 第一层的两个 RBM 参数进行更新，具体更新公式如式(7)－式(10)所示：

$$W_{ij}^{1I}\leftarrow W_{ij}^{1I}+\varepsilon\cdot(h_j^{1I}-h_j^{1T})\cdot\sigma'(h_j^{1I})\cdot v_i^I \tag{7}$$

$$b_j^I\leftarrow b_j^I+\varepsilon\cdot(h_j^{1I}-h_j^{1T})\cdot\sigma'(h_j^{1I}) \tag{8}$$

$$W_{ij}^{1T}\leftarrow W_{ij}^{1T}+\varepsilon\cdot(h_j^{1I}-h_j^{1T})\cdot\sigma'(h_j^{1T})\cdot v_i^T \tag{9}$$

$$b_j^T\leftarrow b_j^T+\varepsilon\cdot(h_j^{1I}-h_j^{1T})\cdot\sigma'(h_j^{1T}) \tag{10}$$

其中，参数的上角标 $I$ 和 $T$ 分别表示这个参数属于 GDBM 和文本模态的 DBM，$\sigma(\ )$ 表示 Logistic 激活函数，$\sigma'(\ )$ 表示其导函数。

CRFMS 模型第二层的预训练过程如图 2 中预训练部分的上半部分所示，其 CD 算法过程与一般的 CD 算法过程仍然有区别[14]。如图 2 中 CRFMS 的预训练部分的第二层所示，两个 RBM 的隐层单元数量增加一倍，通过这样的变化来弥补缺少的自下而上的传递，从而确定可见层单元的状态。因此，依据训练过程的改变，CRFMS 模型第二层的可视单元和隐藏单元的条件概率分布如式(11)和式(12)所示：

$$p(h_j^1=1|h^2)=\sigma(\sum_k W_{jk}^2 h_k^2+\sum_k W_{jk}^2 h_k^2) \tag{11}$$

$$p(h_k^2=1|h^1)=\sigma(\sum_j W_{jk}^2 h_j^1) \tag{12}$$

经过 CRFMS 模型的映射，不同模态的数据被映射到了统一的表示空间，已失去特定模态数据的特殊性，即各自模态数据的表示向量已经具备一定的泛化能力。因此，在第二层的目标函数中，类似式(3)中的超参数 $\alpha$ 和 $\beta$ 已经没有存在的必要性。结合上述说明，并基于式(11)和式(12)中的条件概率分布，第二层的目标函数如式(13)所示：

$$\min L^2=L_D^2+\theta\cdot(L_I^2+L_T^2) \tag{13}$$

CRFMS 模型第二层的预训练过程如下：首先使用前面提到的 CD 算法来对 CRFMS 第二层的两个标准二值 RBM 进行训练，更新两个 RBM 的第一层隐层偏置项、第二层隐层偏置项和它们之间的权重值；然后使用关联误差再次对第二层的两个标准二值 RBM 进行更新，具体更新公式如式(14)－式(17)所示：

$$W_{jk}^{2I}\leftarrow W_{jk}^{2I}+\varepsilon\cdot(h_k^{2I}-h_k^{2T})\cdot\sigma'(h_k^{2I})\cdot h_j^{1I} \tag{14}$$

$$c_k^I\leftarrow c_k^I+\varepsilon\cdot(h_k^{2I}-h_k^{2T})\cdot\sigma'(h_k^{2I}) \tag{15}$$

$$W_{jk}^{2T}\leftarrow W_{jk}^{2T}+\varepsilon\cdot(h_k^{2I}-h_k^{2T})\cdot\sigma'(h_k^{2T})\cdot h_j^{1T} \tag{16}$$

$$c_k^T\leftarrow c_k^T+\varepsilon\cdot(h_k^{2I}-h_k^{2T})\cdot\sigma'(h_k^{2T}) \tag{17}$$

逐层预训练的目的是：通过这一阶段的逐层预训练，将跨模态异构关联学习融入到了训练的过程中，为下一阶段的训练提供了一个较好的初始化权重。

(2)单向融合阶段

将 CRFMS 模型中的图像模态 GDBM 和文本模态 DBM 作为一个整体进行参数的微调，其目标函数如式(18)所示：

$$\min L=L_D+L_I+L_T \tag{18}$$

其中：

$$L_D=L_D^1+L_D^2 \tag{19}$$

$$L_I=\sum_{i=1}\cdot\log p(v_i^I) \tag{20}$$

$$L_T=\sum_{i=1}\cdot\log p(v_i^T) \tag{21}$$

其中，$L_D$，$L_I$ 和 $L_T$ 分别表示 CRFMS 模型中各个层次的关联误差之和、图像数据的似然和文本数据的对数似然。使用关联误差的更新式(9)－式(10)、式(16)－式(17)更新文本模态 DBM 的参数值，图像模态 DBM 的参数值则固定不变。$L_I$ 和 $L_T$ 则通过各自 DBM 微调阶段的变分推断和基于 MCMC 策略的随机逼近来进行参数的更新[14]。

在单向融合阶段，训练策略采用的是轮流优化的思想，即：

1)将预训练阶段得到的两个 DBM 的各项参数作为此阶段两个对应 DBM 的初始化参数；

2)固定文本模态 DBM 的参数，使用变分推断以及基于 MCMC 策略的随机逼近对图像模态 GDBM 进行参数更新；

3)固定图像模态 GDBM 的参数，使用变分推断和基于 MCMC 策略的随机逼近以及关联误差对文本模态 DBM 进行参数更新。

单向融合阶段训练的目的是：经过这一阶段的参数微调，使模型构建起更为复杂的关联关系，即从文本的某个抽象层出发，同时对应多个图像数据抽象层次的关联关系。

CRFMS 模型的整体算法流程如算法 1 所示。

**算法 1**　训练 CRFMS 算法

输入：训练样本；两个 DBM 的 $h^1$ 和 $h^2$ 隐层维度；学习率 $\varepsilon$；超参数 $\alpha$，$\beta$，$\theta$；

输出：模型的所有参数；

预训练阶段：

初始化：初始化所有神经元的偏置项和连接权重值。

CRFMS 第一层：

Repeat：

1. 使用 CD 算法训练 GRBM，更新参数权重参数 $W^{1I}$、显示层偏置项 $a^I$、隐层偏置项 $b^I$；
2. 使用 CD 算法训练 RSRBM，更新参数权重参数 $W^{1T}$、显示层偏置项 $a^T$、隐层偏置项 $b^T$；
3. 使用关联关系再次对权重参数 $W^{1I}$ 和 $W^{1T}$、隐层偏置项 $b^I$ 和 $b^T$ 参数进行更新。

until converge

CRFMS 第二层：

Repeat

1. 使用 CD 算法训练文本模态的 RBM，更新参数权重参数 $W^{2I}$、隐层偏置项 $b^I$、隐层偏置项 $c^I$；
2. 使用 CD 算法训练图像模态的 RBM，更新参数权重参数 $W^{2T}$、隐层偏置项 $b^T$、隐层偏置项 $c^T$；
3. 使用关联关系再次对权重参数 $W^{2I}$ 和 $W^{2T}$、隐层偏置项 $c^I$ 和 $c^T$ 参数进行更新。

until converge

单向融合阶段：

初始化：预训练阶段得到的 CRFMS 模型的各项参数作为单向融合阶段的初始化参数；

Repeat

1. 固定文本模态 DBM 的参数，使用变分推断以及基于 MCMC 策略的随机逼近对图像模态 GDBM 的 $W^{1I}$，$W^{2I}$，$b^I$ 和 $c^I$ 参数进行更新；
2. 固定图像模态 GDBM 的参数，使用变分推断以及基于 MCMC 策略的随机逼近对文本模态 DBM 的 $W^{1T}$，$W^{2T}$，$b^T$ 和 $c^T$ 参数进行更新；
3. 使用关联关系对文本模态 DBM 的权重参数 $W^{1T}$ 和 $W^{2T}$、隐层偏置项 $b^T$ 和 $c^T$ 参数进行更新。

until converge

训练完 CRFMS 模型之后，进行跨模态检索。首先，利用训练好的 CRFMS 模型对测试集中图像和文本模态的数据进行处理，使得不同模态的数据映射到统一的表示空间中；然后，利用相似度匹配来完成检索的工作。如果进行的是以文检图的工作，则给定经过 CRFMS 处理之后的某个文本数据作为查询，并利用欧氏距离对所有经过 CRFMS 处理之后的图像数据进行相似度匹配，最终得到一个按距离递增排列的检索结果列表。以图检文的过程与以文检图的过程类似，这里不再进行赘述。

## 3 实验

### 3.1 实验环境

本文中的所有实验均是在一个拥有主频 2.00 GHz 的 quad E5-2620 CPUs 和 4GB 内存的工作站上完成的；操作系统为 Ubuntu 16.04 64 位；开发环境为 Python 2.7 和 Tensorflow 0.11，开发工具为 PyCharm。

### 3.2 数据集与特征提取

本文所提出的 CRFMS 模型分别在 Wikipedia 数据集、NUS-WIDE-10k 数据集和 Pascal 数据集 3 个公开的真实数据集上进行了实验，这 3 个数据集都是跨模态检索领域中最常使用的数据集，并且它们具有不同的特性，集中体现在样本数据范围不等、所划分的语义类别不等以及每个数据中的文本内容类型不同，分别是篇章、标签和句子。在这 3 个数据集上进行实验能够体现 CRFMS 模型在不同情况下的适应性。本文对数据的预处理方式与文献[2]相同。

(1)Wikipedia[16]。该数据集包含 2866 个图像文本对，划分为 10 个语义类别，随机选取 2173 个图像文本对作为训练集，剩余 693 个图像文本对作为测试集。其中，针对图像部分，提取图像的 1000 维的 PHOW 特征、512 维的 Gist 特征、784 维的 MPEG-7 特征，最终每幅图片表示为一个 2296 维的向量。针对文本数据，使用词袋模型将每个本文表示为一个 3 000 维的向量。

(2)NUS-WIDE-10k[17]。该数据集包含 10 000 个图像文本对，并随机划分其中 8000 个图像文本对作为训练集，另外 2000 个图像文本对作为测试集。针对图像部分，提取每幅图片的 64 维颜色直方图特征、144 维的颜色相关图特征、73 维的边缘方向直方图特征、128 维的小波纹理特征、225 维的块颜色矩特征和 500 维的 SIFT 描述子产生的词袋特征，最终每幅图片表示为 1134 维的向量；针对文本数据，使用词袋模型将每个文本表示为 1000 维的向量。

(3)Pascal[18]。该数据集包含 1000 个图像文本对，分属 20 个语义类别，随机划分其中 800 个图像文本对作为训练集，另外 200 个图像文本对作为测试集。针对图像部分，其处理方式与 Wikipedia 数据集相同。在文本数据部分，使用词袋模型将每个文本表示为一个 1000 维的向量。

### 3.3 模型结构与参数设定

CRFMS 模型在 Wikipedia 数据集、NUS-WIDE-10k 数据集和 Pascal 数据集中，其第二层的表示维度，即 $h^{1I}(h^{1T})$ 的维度分别为 256 维、128 维和 256 维。可以看出，在 NUS-WIDE-10k 数据集中的表示维度小于其余两个数据集的维度，原因是 NUS-WIDE-10k 中初始图像表示维度和初始文本表示维度分别为 1134 维和 1000 维，小于其余两个数据集的初始表示维度。结合文献[3-4]中模型最终表示维度的设定，CRFMS 模型的最终表示维度，即 $h^{2I}(h^{2T})$ 的维度分别设定为 16 维、24 维和 32 维，其目的有两个：1)分别在不同的最终表示维度设定条件下进行跨模态检索的实验，以证明模型最终表示效果的有效性和稳定性；2)3 个维度值的设定都较低，可以有效提升跨模态检索时的效率。

参照文献[3]中对超参数的设定方法，本文通过网格搜索的方法确定了 CRFMS 模型所涉及到的超参数。通过网格搜索的方法，本文最终将 $\alpha$ 值设定为 0.1，$\beta$ 值设定为 $\alpha$ 值的 100 倍，$\theta$ 值则设定为 0.1。

### 3.4 评价指标

(1)平均准确率(mAP)。mAP 是为解决召回率(Recall Rate)、准确率(Precision)和 F-measure 的单点值局限性而提出的一个能够反映全局性能的指标，同时考虑了检索效果的排名情况，是最常用的衡量信息检索结果优劣的标准。

给定一个查询，返回前 $R$ 个结果，其 $mAP$ 的定义如式(22)所示：

$$mAP=\int_0^1 P(R)\mathrm{d}R \tag{22}$$

本文中，返回检索结果数量 $R$ 被设定为 50。

(2)PR 曲线。PR 曲线是以召回率(Recall)为横坐标、以准确率(Precision)为纵坐标绘制得到的曲线。在 PR 曲线图中，PR 曲线与坐标轴所围成的面积越大，这条 PR 曲线所代表模型的检索效果就越好。

### 3.5　对比方法

(1)RBM+CCA[19]。第一阶段使用单模态的 RBM 来建模图像和文本数据，第二阶段使用 CCA 建立多模态数据之间的联系，该模型将表示学习和关联学习分别放在两个阶段进行。

(2)多模态深度玻尔兹曼机(Multimodel DBM)[9]。它使用 GDBM 和包含 RSRBM 的 DBM 建模图像数据和文本数据，再在顶层学习出一个共享表示层，将表示学习和关联学习分为两个阶段来进行。

(3)跨模态多种深度网络模型(Cross-media Multiple Deep Network，CMDN)[6]。该模型融合了包括 DBN，SAE 和 RBM 等在内的多种深度网络，并通过基于共享表示层的多模态关联学习来挖掘复杂的多模态数据之间的内在联系。

(4)对应自编码器(Corr-AE)[2]。Corr-AE 模型可以分为 3 层：第一层使用 GRBM 和 RSRBM 分别对图像和文本数据进行建模；后两层分别使用标准二值 RBM 和 AE 对第一层的表示结果学习更高层次的表示，并在第三层的两个 AE 相对应的表示层之间增加包含相似性约束的对应层次的关联关系。

(5)栈式对应受限玻尔兹曼机(Stacked Corr-RBMs)[3]。Stacked Corr-RBMs 可以分为两层：第一层使用 GRBM 建模图像数据，使用 RSRBM 建模文本数据，然后在两个 RBM 相对应的隐藏层之间增加包含相似性约束的对应层次的关联关系；第二层使用标准二值 RBM 重复第一部分的步骤。

(6)深度关联网络(Deep Correlation Networks，DCN)[5]。DCN 可以分为两层：第一层使用 GRBM 和 RSRBM 建模图像数据和文本数据，然后在两个 RBM 相对应的隐藏层之间增加包含相似性约束的对应层次的关联关系；第二层使用两个 AE 再次处理第一阶段处理之后的结果，并在两个 AE 相对应的表示层之间增加包含相似性约束的对应层次的关联关系。

### 3.6　实验结果及分析

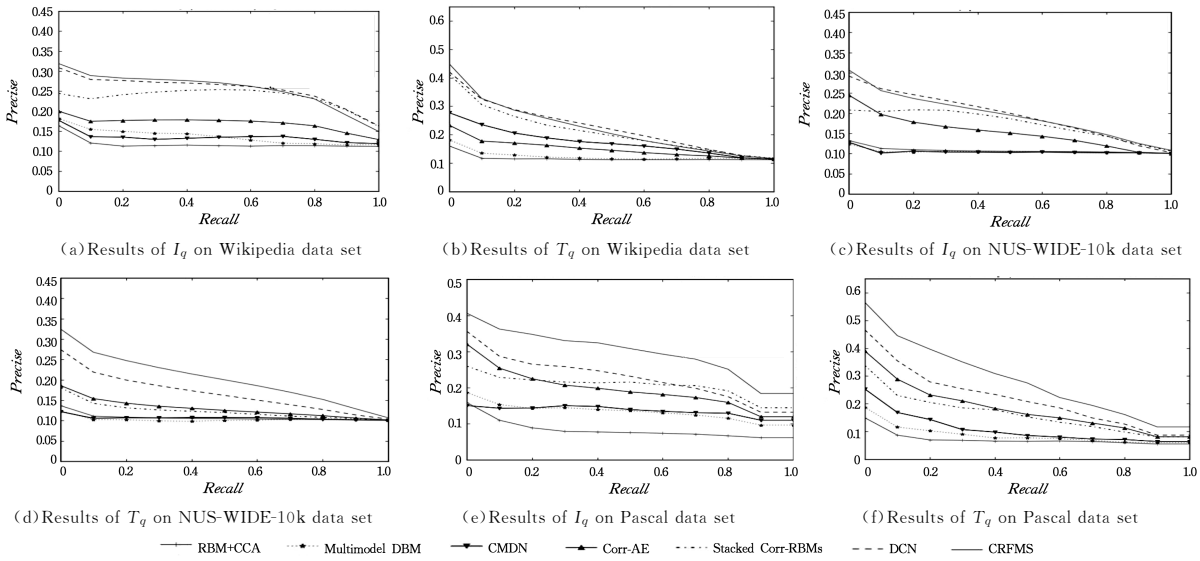
在不同数据集的不同维度条件下的实验结果如表 1 所列。其中，$I_q$ 表示以图检文，$T_q$ 表示以文检图，$Ave$ 表示前面两者 $mAP$ 的平均值。可以看出，本文所提出的 CRFMS 模型的 $mAP$ 值全面优于对比方法。在 Wikipedia 数据集、NUS-WIDE-10k 数据集和 Pascal 数据集下，CRFMS 相比对比方法中表现最好的模型，在 16 维、24 维和 32 维的设定条件下，其平均 $mAP$ 值分别提升了 3.5%，7.6%和 23.7%，尤其是在划分了 20 个语义类别的 Pascal 数据集中，提升效果最为明显。同时，在降低了最终的表示维度之后，许多模型的 $mAP$ 值对比其在文献[2-3,6]中较高维度时的表现，均有较大幅度的下降；而本文所提出的 CRFMS 模型的 $mAP$ 值在不同的最终表示维度下始终稳定在较高的水平，这均体现了基于本文思想所构造的模型的有效性。

表 1　将所有模型的最终表示维度设定为 16 维、24 维和 32 维时各个模型跨模态检索时的 $mAP$ 值

Table 1　$mAP$ scores of all models when final representation dimensions are set to 16-D，24-D and 32-D

| 数据集 | 模型 | 16 维 | | | 24 维 | | | 32 维 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $I_q$ | $T_q$ | $Ave$ | $I_q$ | $T_q$ | $Ave$ | $I_q$ | $T_q$ | $Ave$ |
| Wikipedia | RBM+CCA | 0.170 | 0.179 | 0.175 | 0.165 | 0.169 | 0.167 | 0.177 | 0.168 | 0.173 |
| | Multimodel DBM | 0.161 | 0.181 | 0.171 | 0.178 | 0.186 | 0.182 | 0.162 | 0.206 | 0.184 |
| | CMDN | 0.195 | 0.283 | 0.239 | 0.178 | 0.274 | 0.226 | 0.191 | 0.267 | 0.229 |
| | Corr-AE | 0.201 | 0.281 | 0.241 | 0.207 | 0.286 | 0.247 | 0.225 | 0.239 | 0.232 |
| | Stacked Corr-RBMs | 0.274 | 0.347 | 0.311 | 0.273 | 0.361 | 0.317 | 0.271 | 0.371 | 0.321 |
| | DCN | 0.295 | 0.389 | 0.342 | 0.300 | 0.382 | 0.341 | 0.295 | 0.352 | 0.324 |
| | CRFMS | **0.304** | **0.387** | **0.346** | **0.303** | **0.392** | **0.348** | **0.301** | **0.395** | **0.348** |
| NUS-WIDE-10k | RBM+CCA | 0.165 | 0.175 | 0.170 | 0.180 | 0.186 | 0.183 | 0.174 | 0.180 | 0.177 |
| | Multimodel DBM | 0.149 | 0.199 | 0.174 | 0.178 | 0.187 | 0.183 | 0.182 | 0.172 | 0.177 |
| | CMDN | 0.191 | 0.165 | 0.178 | 0.195 | 0.179 | 0.187 | 0.193 | 0.225 | 0.209 |
| | Corr-AE | 0.241 | 0.224 | 0.233 | 0.264 | 0.217 | 0.241 | 0.256 | 0.224 | 0.240 |
| | Stacked Corr-RBMs | 0.265 | 0.263 | 0.264 | 0.251 | 0.249 | 0.250 | 0.273 | 0.266 | 0.270 |
| | DCN | 0.303 | 0.301 | 0.302 | 0.319 | 0.324 | 0.322 | 0.328 | 0.317 | 0.323 |
| | CRFMS | **0.329** | **0.340** | **0.335** | **0.324** | **0.343** | **0.334** | **0.334** | **0.363** | **0.349** |
| Pascal | RBM+CCA | 0.112 | 0.134 | 0.123 | 0.140 | 0.118 | 0.129 | 0.139 | 0.110 | 0.125 |
| | Multimodel DBM | 0.151 | 0.125 | 0.138 | 0.146 | 0.148 | 0.147 | 0.136 | 0.151 | 0.144 |
| | CMDN | 0.135 | 0.180 | 0.158 | 0.153 | 0.185 | 0.169 | 0.166 | 0.179 | 0.173 |
| | Corr-AE | 0.208 | 0.224 | 0.216 | 0.229 | 0.263 | 0.246 | 0.208 | 0.214 | 0.211 |
| | Stacked Corr-RBMs | 0.212 | 0.232 | 0.220 | 0.192 | 0.243 | 0.218 | 0.225 | 0.278 | 0.252 |
| | DCN | 0.261 | 0.275 | 0.268 | 0.264 | 0.280 | 0.272 | 0.280 | 0.339 | 0.310 |
| | CRFMS | **0.326** | **0.361** | **0.344** | **0.332** | **0.372** | **0.352** | **0.318** | **0.383** | **0.351** |

从图 3—图 5 所示的 PR 曲线图中可以看出，在不同数据集的不同维度下，所提 CRFMS 模型的 PR 曲线所围成的面积要比所有对比方法所围成的面积大，这从另一个角度证明了所提 CRFMS 模型的有效性，更证明了本文算法思想的合理性。

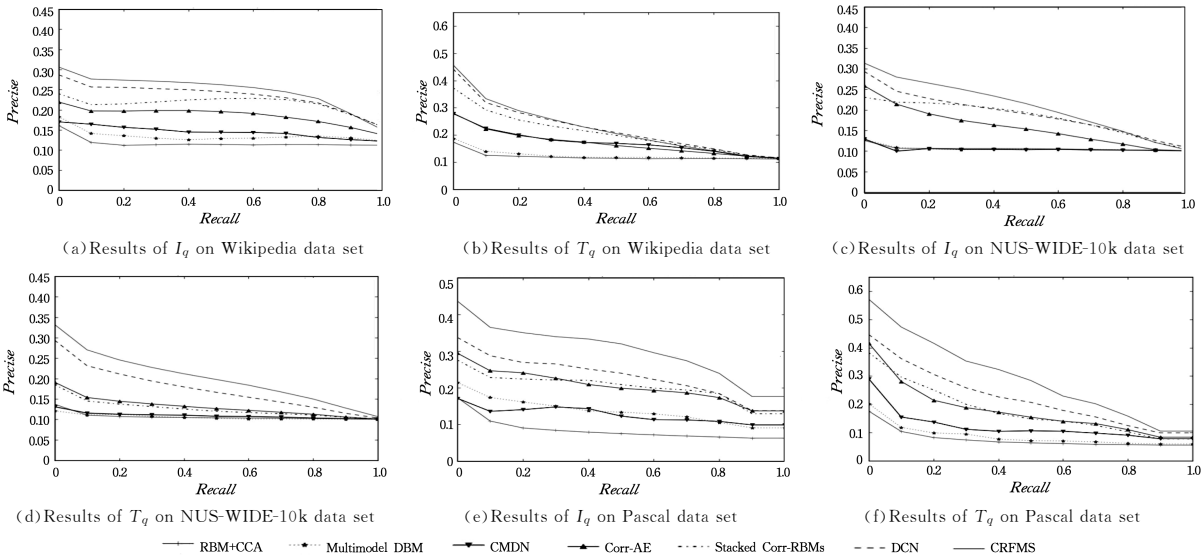



(a)Results of $I_q$ on Wikipedia data set (b)Results of $T_q$ on Wikipedia data set (c)Results of $I_q$ on NUS-WIDE-10k data set

(d)Results of $T_q$ on NUS-WIDE-10k data set (e)Results of $I_q$ on Pascal data set (f)Results of $T_q$ on Pascal data set

图 3 将所有模型的最终表示维度设定为 16 维时,各个模型在 3 个数据集上的 PR 曲线

Fig. 3 PR curves of all models on three datasets when final representation dimension is set to 16-D

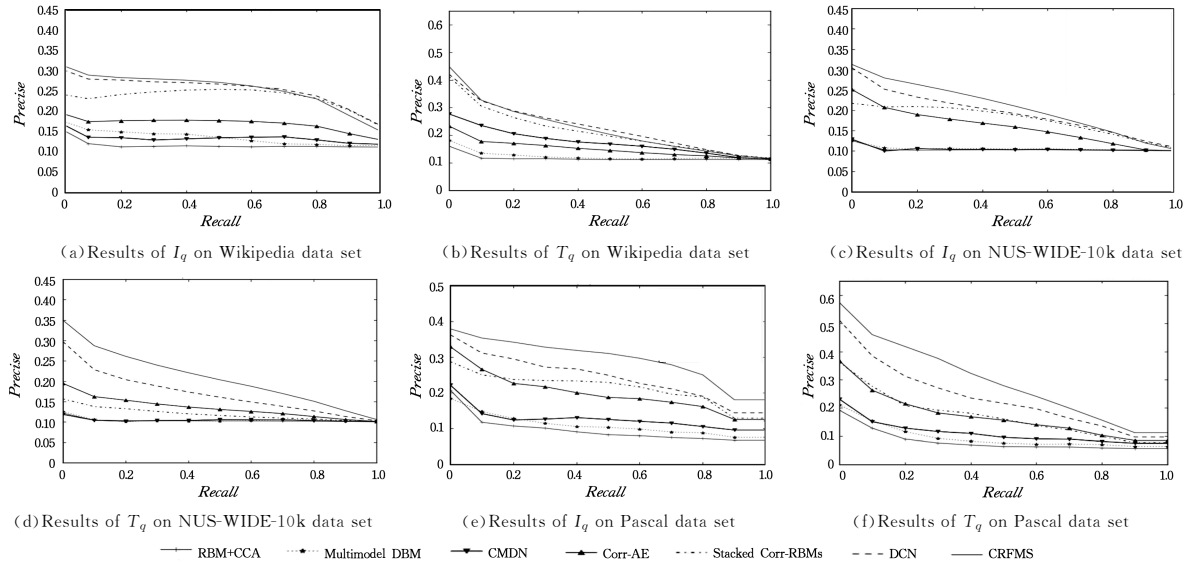



(a)Results of $I_q$ on Wikipedia data set (b)Results of $T_q$ on Wikipedia data set (c)Results of $I_q$ on NUS-WIDE-10k data set

(d)Results of $T_q$ on NUS-WIDE-10k data set (e)Results of $I_q$ on Pascal data set (f)Results of $T_q$ on Pascal data set

图 4 将所有模型的最终表示维度设定为 24 维时,各个模型在 3 个数据集上的 PR 曲线

Fig. 4 PR curves of all models on three datasets when final representation dimension is set to 24-D




(a)Results of $I_q$ on Wikipedia data set (b)Results of $T_q$ on Wikipedia data set (c)Results of $I_q$ on NUS-WIDE-10k data set

(d)Results of $T_q$ on NUS-WIDE-10k data set (e)Results of $I_q$ on Pascal data set (f)Results of $T_q$ on Pascal data set

图 5 将所有模型的最终表示维度设定为 32 维时,各个模型在 3 个数据集上的 PR 曲线

Fig. 5 PR curves of all models on three datasets when final representation dimension is set to 32-D

**结束语**　本文通过将 DBM 与包含相似性约束的关联学习进行有机的结合，借助了 DBM 无向图的特性，在异构模态数据之间构造了更为复杂的关联关系，即文本模态数据的某个抽象层次同时与多个图像模态数据的抽象层次构建关联关系，打破了之前不同模态数据之间只包含 1-1 的对应关联关系或只存在共享表示层的局限，从而挖掘出了更细粒度的异构多模态数据的潜在语义关联。实验证明，本文所建立的 CRFMS 模型可以有效地提升跨模态检索的精确度，说明 CRFMS 模型具有合理性和有效性。

近年来，深度学习中的注意力模型（Attention Model，AM）被广泛应用于自然语言处理、图像识别及语音识别等各种不同类型的深度学习任务中，是深度学习技术中最值得关注与深入了解的核心技术之一。在未来的工作中，如何将 AM 整合到跨模态检索模型之中，使得不同模态数据在进行关联学习时变得更有侧重点，是未来工作中需要研究的内容。

## 参 考 文 献


[1] FENG F X. Deep learning for cross-modal retrieval[D]. Beijing: Beijing University of Posts and Telecommunications, 2015. (in Chinese)
冯方向. 基于深度学习的跨模态检索研究[D]. 北京：北京邮电大学，2015.

[2] FENG F, WANG X, LI R. Cross-modal retrieval with correspondence autoencoder[C]// Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 7-16.

[3] FENG F, LI R, WANG X. Deep correspondence restricted Boltzmann machine for cross-modal retrieval[J]. Neurocomputing, 2015, 154: 50-60.

[4] WANG W, OOI B C, YANG X, et al. Effective multi-modal retrieval based on stacked auto-encoders[J]. Proceedings of the VLDB Endowment, 2014, 7(8): 649-660.

[5] CAI G, FENG Y, LIN Q. Cross-modal retrieval based on deep correlated network[C]// 2017 3rd IEEE International Conference on Computer and Communications (ICCC). IEEE, 2017: 1226-1231.

[6] PENG Y, HUANG X, QI J. Cross-media Shared Representation by Hierarchical Learning with Multiple Deep Networks[C]// International Joint Conference on Artificial Intelligence (IJCAI). IEEE, 2016: 3846-3853.

[7] WANG K, YIN Q, WANG W, et al. A comprehensive survey on cross-modal retrieval[J]. arXiv preprint arXiv: 1607. 06215, 2016.

[8] SALAKHUTDINOV R, HINTON G. Deep boltzmann machines[C]// Artificial Intelligence and Statistics. IEEE, 2009: 448-455.

[9] SRIVASTAVA N, SALAKHUTDINOV R R. Multimodal learning with deep boltzmann machines[C]// Advances in Neural Information Processing Systems. 2012: 2222-2230.

[10] CHO K H, RAIKO T, ILIN A. Gaussian-bernoulli deep boltzmann machine[C]// The 2013 International Joint Conference on Neural Networks (IJCNN). IEEE, 2013: 1-7.

[11] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[R]. Technical Teport, University of Toronto, 2009.

[12] WELLING M, ROSEN-ZVI M, HINTON G E. Exponential family harmoniums with an application to information retrieval[C]// Advances in Neural Information Processing Systems. 2005: 1481-1488.

[13] HINTON G E, SALAKHUTDINOV R R. Replicated softmax: an undirected topic model[C]// Advances in Neural Information Processing Systems. 2009: 1607-1614.

[14] SALAKHUTDINOV R, LAROCHELLE H. Efficient learning of deep Boltzmann machines[C]// Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. 2010: 693-700.

[15] HINTON G E. Training products of experts by minimizing contrastive divergence[J]. Neural Computation, 2002, 14(8): 1771-1800.

[16] RASIWASIA N, COSTA PEREIRA J, COVIELLO E, et al. A new approach to cross-modal multimedia retrieval[C]// Proceedings of the 18th ACM International Conference on Multimedia. ACM, 2010: 251-260.

[17] CHUA T S, TANG J, HONG R, et al. NUS-WIDE: a real-world web image database from National University of Singapore[C]// Proceedings of the ACM International Conference on Image and Video Retrieval. ACM, 2009.

[18] FARHADI A, HEJRATI M, SADEGHI M, et al. Every picture tells a story: Generating sentences from images[M]// Computer Vision-ECCV 2010. Berlin: Springer, 2010: 15-29.

[19] NGIAM J, KHOSLA A, KIM M, et al. Multimodal deep learning[C]// Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011: 689-696.