第48卷第4期 中国科学技术大学学报 Vol.48,No.4
2018年4月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Apr. 2018

文章编号:0253-2778(2018)04-0322-09

基于深度典型相关性分析的跨媒体语义检索

王 述[1,2],史忠植[1]

(1.中国科学院计算技术研究所智能信息处理重点实验室,北京 100190;2.中国科学院大学工程科学学院,北京 100049)

摘要:基于典型相关性分析的跨媒体检索是一种将不同媒体特征通过相关性分析映射到同构的最大相关子空间,并在子空间中完成跨媒体数据间的相似性比较和检索的方法.典型相关性分析(canonical correlation analysis, CCA)是一种线性模型,并不能很好地挖掘跨媒体数据中的复杂相关关系.为此针对深度典型相关性分析(deep CCA, DCCA)的结构进行改进,使用隐含狄利克雷分布(latent Dirichlet allocation, LDA)发现文本语义信息并学习语义映射,提出了跨媒体深度相关性学习模型(cross-media correlation learning with deep canonical correlation analysis, CMC-DCCA)以及跨媒体语义相关性检索方法(cross-media semantic correlation retrieval, CMSCR).在维基百科文本图像数据集上的实验证明,CMC-DCCA模型能够较好地挖掘跨媒体数据中的复杂相关关系,CMSCR在跨媒体检索中具有较好的性能.

关键词:典型相关性分析;深度典型相关性分析;语义映射;跨媒体检索

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2018.04.008

引用格式:王述,史忠植.基于深度典型相关性分析的跨媒体语义检索[J].中国科学技术大学学报,2018,48(4):322-330.

WANG Shu, SHI Zhongzhi. Cross-media semantic retrieval with deep canonical correlation analysis[J]. Journal of University of Science and Technology of China, 2018,48(4):322-330.

Cross-media semantic retrieval with deep canonical correlation analysis

WANG Shu[1,2], SHI Zhongzhi[1]

(1. *Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190;
2. *University of Chinese Academy of Sciences, School of Engineering Science, Beijing* 100049)

Abstract: The cross-media retrieval with canonical correlation analysis (CCA) is a method to map different media features to the largest correlation isomorphism subspace through the canonical correlation analysis, and compare the similarity between cross-media data in the subspace. However CCA is a linear model and can not adequately exploit the complex correlation between cross-media data. The structure of the traditional deep canonical correlation analysis (DCCA) is improved, and the latent dirichlet allocation (LDA) is used to discover the semantic information in the text data and learns the semantic mapping. The cross-media correlation learning with deep canonical correlation analysis (CMC-DCCA) and the cross-media semantic correlation retrieval (CMSCR) are proposed. Experiments on the Wikipedia text image dataset shows that the CMC-DCCA model can mine the complex correlation between cross-media data

收稿日期:2017-06-01;**修回日期**:2017-07-14

基金项目:国家重点基础研究发展计划(973)(2013CB329502),国家自然科学基金(61035003)资助.

作者简介:王述,男,1989年生,硕士生.研究方向:语义检索、数据挖掘.E-mail:wangshu04@baidu.com

通讯作者:史忠植,博士/研究员.E-mail:shizz@ics.ict.ac.cn

better, and that CMSCR has better performance in cross-media retrieval.

Key words: canonical correlation analysis; deep canonical correlation analysis; semantic mapping; cross-media retrieval

0　引言

由于网络上多媒体数据的快速增长,跨媒体检索在人工智能、多媒体信息检索和计算机视觉等领域成为多年来关注及研究的焦点. 跨媒体检索旨在挖掘表达同一语义的不同模态数据之间的相关关系. 本文研究的跨媒体检索主要是文本和图像之间的相互检索.

跨媒体检索的主要挑战是跨媒体异构鸿沟和跨媒体语义鸿沟. 异构鸿沟是指不同媒体数据底层特征之间由于特征维数及属性上的不同导致彼此之间存在异构性问题. 语义鸿沟是指媒体数据底层特征与其高层语义间可能存在不一致问题. 目前,针对这两个难题,已经提出了许多的跨媒体检索方法. 我们将这些方法分为两大类:浅层模型方法和深层模型方法.

浅层模型方法主要目的在于通过直接挖掘不同媒体数据底层特征之间的相关关系进行检索. Rasiwasia 等[1] 提出了一种使用典型相关性分析(CCA)将文本和图像特征映射到最大相关子空间进行相似性度量的方法. 随后,一系列基于 CCA 的扩展方法被提出用于构建跨媒体相关性学习模型[2-6]. Gong 等[7] 首次将单一类别或者多个概念标签表示的高层图像语义作为一个视图并结合二视图的 CCA 提出了三视图 CCA 方法. 张博等[8] 提出的混合概率典型相关性分析模型(MixPCCA)通过混合多个局部线性 PCCA 模型不仅提供了一种捕捉复杂的全局非线性相关性的解决方案,而且还具备检测只在局部区域才存在的相关性的能力. 针对弱匹配多模态数据的相关性建模问题,张博等[9] 提出了一种弱匹配概率典型相关性分析模型(semi-paired probabilistic CCA, SemiPCCA),解决在匹配样本不足的情况下出现的过拟合问题. 虽然这些方法在跨媒体检索中取得了不错的效果,但是在跨媒体检索的潜在空间中,语义上一致的图像和文本应该是彼此互相靠近的,而使用 CCA 及其扩展方法仅仅最大化图像及对应文本之间的相关性,并不能满足这个要求.

深层模型方法采用多层非线性处理单元组成的网络对跨媒体数据进行特征提取,并具有比浅层模型更强大的表示能力. Srivastava 等[10] 提出使用深度置信网络(deep belief network,DBN)来学习图像和文本输入的联合空间的生成模型. 基于 DBN 的思想,Feng 等[11] 提出了通过对应自动编码机(corres-pondence auto-encoder,Corr-AE)来学习图像和文本输入的潜在空间的跨模态检索方法. 该模型通过对两个单模自动编码器的隐藏表示进行关联来构造. 它通过对不同模态数据学习自编码表达,并让不同模态在隐含空间的表达相似来学习跨模态之间的映射函数. Wang 等[12] 提出了一种用于跨模态语义映射的正则化的深层神经网络(regularized deep neural network, RE-DNN). 其设计和实现了 5 层神经网络,用于将视觉和文本特征映射到公共语义空间,从而可以度量不同模态数据之间的相关性. Andrew 等[13] 提出了深度典型相关性分析(deep canonical correlation analysis, DCCA). DCCA 是一种学习不同模态数据形式间的复杂非线性投影,使得所得到的数据表示是高度线性相关的深度学习方法.由于训练数据不足的原因,这些深度方法往往会出现过拟合问题. 与浅层方法相似的是,深层方法也无法解释潜在空间中距离相近的文本和图像为什么是相似的.

本文的主要贡献如下:

(Ⅰ)针对 DCCA 进行结构改进,构建一个跨媒体相关性学习模型. 其中线性投影层的训练是同非线性隐含层的训练结合在一起的,以确保得到相对较少并且更加抽象和更加准确的特征表示,从而能较好地解决传统 DCCA 的过拟合问题;

(Ⅱ)提出一种语义对齐的数据驱动方法,根据多媒体数据中发现的词汇概念及其标签信息,分别对图像和文本数据训练语义映射实现跨媒体检索;

(Ⅲ)在维基百科文本图像数据集的实验结果表明,本文提出的跨媒体语义检索算法(CMSCR)是有效的.

1　相关工作

1.1　深度相关性分析

CCA 是一种线性数学模型,是用于多数据源的

多视图和多尺度分析的一种经典统计方法，其主要目的为找到具有最大相关性的不同类型数据的线性投影. Andrew 等在 2013 年提出了 DCCA，DCCA 是将深层网络与 CCA 进行结合以学习灵活的非线性表示. DCCA 将两个视图通过多个非线性变换的全连接层来计算两个视图的表示，其结构如图 1 所示，其包括两个深度网络，通过学习使输出层是最大相关的，其中下排节点表示输入的特征，中间 3 排节点表示隐藏层，上排节点表示输出层，每个网络据均有 d 层.

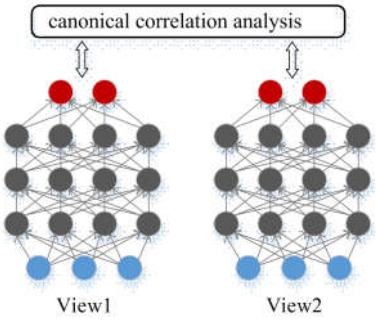

图 1 深度典型相关性分析结构示意图

Fig.1 The schematic of deep CCA

假设用于 View 1 的网络中的每个中间层具有 c_1 个单元，并且最后(输出)层具有 o 个单元. 设 $x_1 \in \mathbb{R}^{n_1}$ 为第一视图中的一个实例. 实例 x_1 通过第 1 层的输出为 $h_1 = s(W_1^1 x_1 + b_1^1) \in \mathbb{R}^{c_1}$，其中 $W_1^1 \in \mathbb{R}^{c_1 \times n_1}$ 为权值矩阵，$b_1^1 \in \mathbb{R}^{c_1}$ 为贝叶斯向量，s：$\mathbb{R} \mapsto \mathbb{R}$为一个分量应用的非线性函数. 然后可以使用输出值 h_1 计算下一层的输出 $h_2 = s(W_2^1 h_1 + b_2^1) \in \mathbb{R}^{c_1}$，以此类推，可以计算出一个具有 d 层的网络的最后的表示为 $f_1(x_1) = s(W_d^1 h_{d-1} + b_d^1) \in \mathbb{R}^o$. 给定 View 2 的一个实例 x_2，可以通过相同的方法计算出用不同的参数 W_l^2 和 b_l^2 得到的最终表示 $f_2(x_2)$. 目标是联合学习两个视图的参数 W_l^v 和 b_l^v 使得 $\mathrm{corr}(f_1(x_1), f_1(x_1))$ 的值尽可能大. 假定 θ_1 是 View 1 的所有参数 W_l^1 和 b_l^1 的向量，其中 $l = 1, 2, \cdots, d$. 同样 θ_2 与此类似，可得

$$(\theta_1^*, \theta_2^*) = \underset{(\theta_1, \theta_2)}{\mathrm{argmax}} \mathrm{corr}(f_1(X_1; \theta_1), f_2(X_2; \theta_2)) \tag{1}$$

为了得到 (θ_1^*, θ_2^*)，需要在训练数据上计算相关目标的梯度.

设在一个大小为 m 的训练集上矩阵 $\boldsymbol{H}_1 \in \mathbb{R}^{o \times m}$，$\boldsymbol{H}_2 \in \mathbb{R}^{o \times m}$ 其列为由两个视图上的深层模型产生的最上层表示结果. 设 $\overline{\boldsymbol{H}}_1 = \boldsymbol{H}_1 - \frac{1}{m}\boldsymbol{H}_1 \boldsymbol{1}$，$\overline{\boldsymbol{H}}_2 = \boldsymbol{H}_2 - \frac{1}{m}\boldsymbol{H}_2 \boldsymbol{1}$ 为中心值矩阵，并且定义 $\hat{\Sigma}_{12} = \frac{1}{m-1}\overline{\boldsymbol{H}}_1 \cdot \overline{\boldsymbol{H}}'_2$，$\hat{\Sigma}_{11} = \frac{1}{m-1}\overline{\boldsymbol{H}}_1 \overline{\boldsymbol{H}}'_1 + r_1 I$ 和 $\hat{\Sigma}_{22} = \frac{1}{m-1}\overline{\boldsymbol{H}}_2 \cdot \overline{\boldsymbol{H}}'_2 + r_2 \boldsymbol{I}$，其中 r_1，r_2 为正则化项. 假设 $r_1 > 0$，$r_2 > 0$ 从而使得 $\hat{\Sigma}_{11}$，$\hat{\Sigma}_{22}$ 是正定的.

$\boldsymbol{H}_1$ 和 $\boldsymbol{H}_2$ 的前 k 个分量的总相关性是矩阵 $\boldsymbol{T} = \hat{\Sigma}_{11}^{-1/2} \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1/2}$ 的前 k 个奇异值的和. 令 $k = o$，则这正是矩阵 $\boldsymbol{T}$ 的迹范数.

$$\mathrm{corr}(\boldsymbol{H}_1, \boldsymbol{H}_2) = \|\boldsymbol{T}\|_{\mathrm{tr}} = \mathrm{tr}\,(\boldsymbol{T}'\boldsymbol{T})^{1/2} \tag{2}$$

为了计算 $\mathrm{corr}(\boldsymbol{H}_1, \boldsymbol{H}_2)$ 的梯度，可以分别计算其相对于 $\boldsymbol{H}_1$ 和 $\boldsymbol{H}_2$ 的梯度，然后使用反向传播的方法求解. 设矩阵 $\boldsymbol{T} = \hat{\Sigma}_{11}^{-1/2} \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1/2}$ 的奇异值分解为 $\boldsymbol{T} = UDV'$，则有

$$\frac{\partial \mathrm{corr}(\boldsymbol{H}_1, \boldsymbol{H}_2)}{\partial \boldsymbol{H}_1} = \frac{1}{m-1}(2 \nabla_{11} \overline{\boldsymbol{H}}_1 + \nabla_{12} \overline{\boldsymbol{H}}_2) \tag{3}$$

其中，

$$\nabla_{12} = \hat{\Sigma}_{11}^{-1/2} UV' \hat{\Sigma}_{22}^{-1/2} \tag{4a}$$

$$\nabla_{11} = -\frac{1}{2}\hat{\Sigma}_{11}^{-1/2} UDU' \hat{\Sigma}_{11}^{-1/2} \tag{4b}$$

同样地，能得到 $\frac{\partial \mathrm{corr}(\boldsymbol{H}_1, \boldsymbol{H}_2)}{\partial \boldsymbol{H}_2}$ 与之对称的表达式.

1.2 隐含狄利克雷分布

主题模型是一种用于文本处理与数据挖掘的非常重要的方法，它可以有效地从文本语义中提取主题信息. 随着主题模型在文本分析领域被广泛应用，并且在图像分类领域也取得十分成功的效果，许多研究者开始提出用主题模型对不同类型媒体数据进行主题建模从而挖掘不同媒体数据之间在语义主题上的相关关系. Blei 等[14]在 2003 年提出的隐含狄利克雷分布(latent Firichlet allocation，LDA)是一种被广泛使用的概率主题模型.

LDA 中，每个文档被视为由各种主题混合而成. 它可以将文档集中每篇文档的主题按照概率分布的形式给出. 文档中每个单词以一定概率选择了某个主题，词汇表中的单词以不同的出现概率构成各主题，形成主题字典(主题-单词分布表)，每个单词的主题和单词本身服从多项式分布.

具体来说，以狄利克雷分布为先验分布 α 对语料库中的每一个文本文档生成文档的多项式主题分

布 $\boldsymbol{\theta}_i$，然后再以狄利克雷分布 β 为先验分布生成主题 k 的多项式主题词汇分布 $\boldsymbol{\varphi}_k$. 对于语料库中的每一篇文本文档 i 中的每一个单词，从文档对应的主题多项式分布 $\boldsymbol{\theta}_i$ 中抽取一个主题 $z_{i,j}$，然后从主题 $z_{i,j}$ 对应的单词多项式分布 $\boldsymbol{\varphi}_{z_{i,j}}$ 中抽取一个单词 $w_{i,j}$，将这个过程重复 N_i，文档 W_i 就生成了. 文档被表示为潜在主题上的随机组合，其中每个主题表示为其对应单词的概率分布.

2　跨媒体深度相关性学习

2.1 改进结构的 DCCA 模型

DCCA 模型的深层网络结构是基于端对端的原则来构建的. 网络的第 1 层为网络的输入层，后续的每一层都与其前一层连接，在最后一层中通过对输出的训练结果与样本真实数据不一致的进行惩罚，得到损失值，并通过最小化损失值来获得网络的最优化参数.

我们对 DCCA 的网络结构进行改进，将原有的深度网络中的第 1 个隐含层改造为线性损失层，线性损失层中的神经元的激活函数为线性函数. 结构如图 2 所示. 新的网络训练的优化目标变为线性损失层的线性投影损失值 L_{linear}，输出层的损失值 L_{out} 和正则化惩罚项的加权和.

$$\text{Loss}=\lambda_1 L_{\text{linear}}+\lambda_2 L_{\text{out}}+\lambda_3 \|\Theta\|_{\text{reg}} \tag{5}$$

在反向传播阶段过程中，线性损失层的误差项将是从最后输出层反向传播得到的误差与其自身输出产生的误差之和.

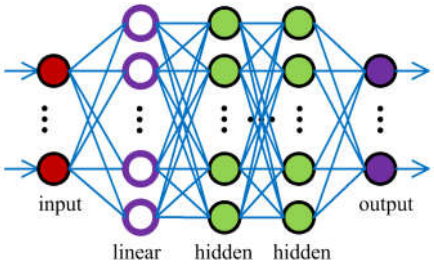

图 2　改进的网络结构示意图

Fig.2　Improved network structure

在改进的网络结构中，线性层通过线性投影完成特征的粗略表示，并将得到的特征表示提供给上层非线性层. 需要强调的是，线性投影层的训练是同非线性投影层的训练结合在一起的，以确保线性投影能够很好地与非线性处理阶段进行匹配. 接下来的非线性层可以逐步将复杂的输入数据表示成相对较少并且更加抽象和更加准确的特征表示.

2.2　跨媒体深度相关性学习模型

本文采用改进结构的 DCCA 构建一个基于深度典型相关性分析的跨媒体相关性学习模型（cross-media correlation learning with deep canonical correlation analysis，CMC-DCCA）. CMC-DCCA 体系结构如图 3 所示. CMC-DCCA 包含两个分支，每个分支都包含一个输入层，若干个隐藏层和一个线性损失层（loss 1）和一个非线性损失层（loss 2），分别负责学习图像和文本的潜在表示. DCCA 试图找到一种映射方法来最大化文本和图像之间的相关性. 与 DCCA 所采用的方法相同，在图 3 所示的 CMC-DCCA 结构中，上下子网络最右端的线性 CCA 层用于对图像和文本视图的潜在表示进行相关性学习.

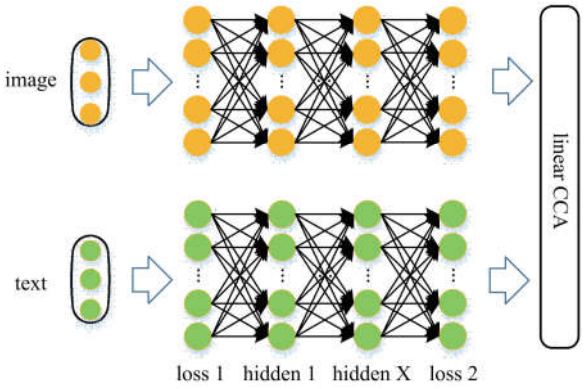

图 3　CMC-DCCA 结构示意图

Fig.3　The structure of CMC-DCCA

假设 I_j 和 T_j 分别表示第 j 个损失层中的图像和文本的潜在表示. 传统的 CCA 模型试图将图像和文本分别投影到 k 维潜在空间中，可以最大化图像和文本之间的相关性. CCA 模型的损失函数如下：

$$L_{\text{CCA}}=c-\text{corr}(I_j,T_j) \tag{6}$$

式中，c 为矩阵 I_j 和 T_j 的相关性上边界. 因此，CMC-DCCA 的整个损失函数表示为

$$\text{Loss}=\alpha L_{\text{CCA}}(I_1,T_1)+L_{\text{CCA}}(I_2,T_2)+\lambda\|\Theta\|_F^2 \tag{7}$$

式中，$L_{\text{CCA}}(I_1,T_1)$ 为线性损失，$L_{\text{CCA}}(I_2,T_2)$ 为非线性损失，λ 为正则化惩罚项系数，α 为深层网络中线性层和非线性输出层之间的层间折衷. 为了优化式(7)，我们需要在训练数据上计算相关目标的梯度.

我们采用文献[13]中的方法对梯度进行推导和计算. 相关性目标函数是整个训练集合的函数，不

能分解为在所有数据点上的和，所以并不清楚如何使用一次一个地对数据点进行操作的随机优化过程. 本文采用 L-BFGS 二阶优化方法[15]来进行最小化优化，其对于优化含有大规模变量的方法效果较好.

2.3 跨媒体相关性计算

训练后，需要对图像文本对的相关性进行计算. 通过相关性学习我们可以得到文本数据的 d 维相关性子空间 $UT \subset \Re^T$，图像数据的 d 维相关性子空间 $UI \subset \Re^I$，UT 和 UI 在文本图像数据上是最大相关的. 在 UT 和 UI 之间存在一个自然可逆的映射关系，d 维子空间中的图像和文本都是一一对应的，可以用 $w_{i,1} \leftrightarrow w_{t,1}, \cdots, w_{i,d} \leftrightarrow w_{t,d}$ 表示，因此两种类型的媒体数据通过相关性学习可得到一种简洁有效的表示. 设定向量 p_T 和 p_I 分别为文本和图像数据在 d 维同构相关性子空间中的坐标. 则在潜在特征空间中的相关性计算如下：

$$\mathrm{sim}(I,T)=d(p_I,p_T) \tag{8}$$

式中，$d(\cdot)$ 为距离度量函数，本文采用中心归一化相关系数作为距离度量函数.

3 跨媒体语义相关性检索

在跨媒体检索的潜在空间中，语义上一致的图像和文本应该是彼此互相靠近的. 由于使用 CCA 及其扩展方法仅仅最大化图像及对应文本之间的相关性，并不能满足这个要求，因此需要挖掘跨媒体数据中不同类型媒体数据间内在的语义信息. 本文提出了一种新的跨媒体检索方法(CMSCR)来解决这个问题，CMSCR 的框架结构如图 4 所示. 其在跨媒体相关性学习之后采用语义对齐的数据驱动方法，根据从多媒体数据中发现的词汇概念及其标签信息分别对图像和文本数据训练语义映射，从而得到图像语义空间及文本语义空间. 图像和文本被表示为相对于同一组语义概念的后验概率向量，所以图像语义空间和文本语义空间是同构的.

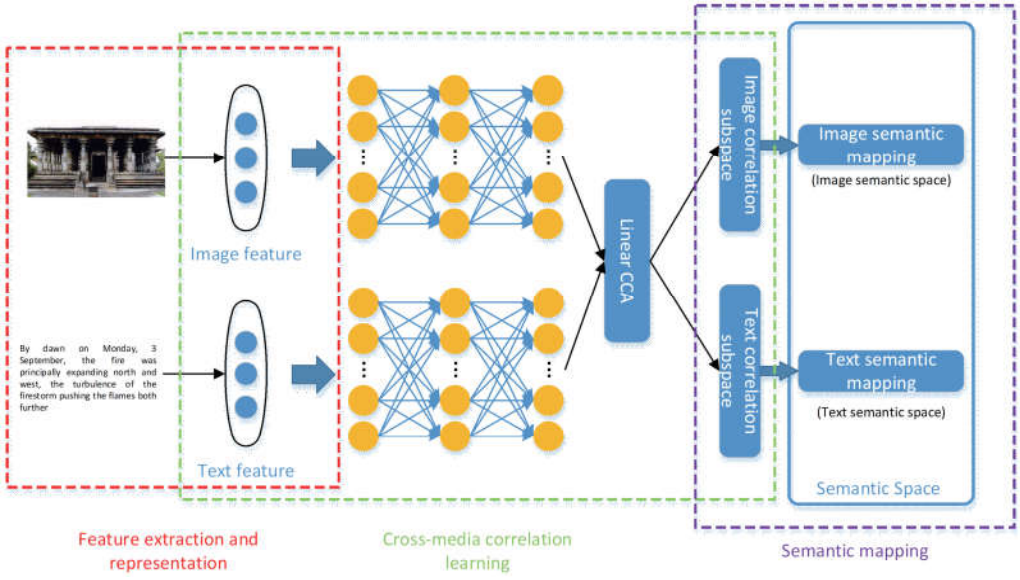

图 4 CMSCR 的框架结构示意图

Fig.4 The structure of CMSCR

为了更加清晰地表述，我们首先介绍自动发现文本数据的语义信息，然后将它们用于训练图像和文本语义映射.

3.1 探索语义信息

在机器学习和自然语言处理中，主题模型是一种用于发现文档集合中的抽象“主题”的统计模型. 主题建模是一种经常使用的文本挖掘方法，用于在文本中发现其所隐含的语义结构. 对于每个文本，主题部分用 0 和 1 之间的概率值来衡量文本与每个主题的相关性.

我们发现主题词和主题概率分布比较适合作为语义词汇概念及文档语义表示. 主题词是对整个文档集合的一个全面总结，因此一个由主题词构成的语义词汇概念表对文档语义具有较高的描述. 此外，通过概率主题模型将相关概念组合成更抽象的概念通常能改进学习的概念分类器的性能. 如从文档中预测“地理”这种比较抽象的概括性概念相较于图像中“公园”，“山脉”和“建筑”等特定概念要更准

确. 这是由于在一定程度上每种概念词的预测准确率是由其在训练集中的数量来决定的.

LDA 模型是一种被广泛使用的概率主题模型. 在 LDA 中,每个文档被视为由各种主题混合而成. 它可以将文档集中每篇文档的主题按照概率分布的形式给出. 我们对训练集中的文本文档使用 LDA 自动构建其语义词汇概念及文档语义表示. 然后将自动发现的主题词汇作为语义词汇概念表,主题概率分布作为文档的语义表示用来训练视觉语义映射和文本语义映射.

3.2　训练语义映射

语义映射的定义为通过预测每个实例与词汇概念的相关性,得到一个从视觉或者文本特征空间到相同语义空间的映射关系. 我们对视觉语义映射和文本语义映射的训练采用类似的方法. 下面首先介绍视觉语义映射的学习算法,然后介绍如何将所提出的算法应用于训练文本语义映射.

3.2.1　训练图像语义映射

我们用 $\varphi_W(x):\mathbb{R}^{D_I}\rightarrow\mathbb{R}^k$ 表示图像语义映射,表示为每个图像样本 $x\in\mathbb{R}^{D_I}$ 在语义空间中的 k 维投影. 视觉语义映射的定义如下:

$$\varphi_W(x)=\boldsymbol{W}^{\mathrm{T}}x \tag{9}$$

式中,$\boldsymbol{W}\in\mathbb{R}^{D_I\times k}$ 是由对应于每个词汇概念分类器 $w_c\in\mathbb{R}^{D_I}$ 的权重向量所构成的图像预测矩阵.

每个词汇概念分类器 w_c 是在训练数据集 $D_c=\{(x_i,y_i),i=1\cdots N\}$ 上进行训练的,其中 $x_i\in\mathbb{R}^{D_I}$ 表示训练集中的每个图像样本. 例如,在基于 Wikipedia 文章的训练数据集上,x_i 是指从文章的图像中提取的视觉特征在跨媒体相关性学习得到的图像最大相关性子空间中的向量表示. 此外,y_i 是表示图像样本 x_i 与概念 c 的相关性的概念标签.

我们使用 LDA 模型来预测文章主题并从中提取概念标签. 对于每个图像样本 x_i,我们使用其对应文本文档的主题概率分布来提取其概念标签 y_i. 主题概率分布将文档主题表示为主题词汇表中每个主题词所代表的介于 0 到 1 之间的连续概率值. 文献[2]直接采用连续概率值作为标签去训练概念分类器. 本文对主题的概率分布中每个主题的概率值进行排序,并且将它们的排序结果作为标签. 换句话说,词汇概念分类器 w_c 被训练用来预测图像样本相对于主题概念的相关性的排序结果.

首先测量每个概念标签的主题概率分布中每个主题的排名来定义每个概念标签 $y_i\in D_c$,计算公式如下:

$$y_i=\sum_{i=1}^{N}\mathbb{I}(\theta_i^c<\theta_j^c) \tag{10}$$

式中,$\mathbb{I}(\cdot)$ 为指标函数,此外 θ_i^c 是样本 i 在主题 c 下所得到的主题概率值. 在确定标签之后,概念分类器可以通过任何排序学习的方法来进行训练. 我们采用 RankSvm 算法[16],通过最小化公式所示的目标函数来学习排序函数,即

$$\frac{\lambda}{2}\|w_c\|^2+\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1,y_i<y_j}^{N}\max(0,1-\boldsymbol{W}_c^{\mathrm{T}}(x_j-x_i)) \tag{11}$$

式中,λ 是正则化惩罚参数.我们对每个主题概念 c 独立使用随机梯度下降的方法来最小化公式(11),需要注意的是,对于每个主题来说文档的排序结果是不同的.

3.2.2　训练文本语义映射

文本语义映射的训练方法与图像语义映射的方法类似,但文本映射训练过程中并不是使用图像视觉特征,而是在文本特征上训练概念分类器.

对于每个词汇概念 c,定义训练数据集 $D_c=\{(x_i,y_i),i=1\cdots N\}$,其中 $x_i\in\mathbb{R}^{D_T}$ 表示训练集中的每个文本样本. 例如,在一组 Wikipedia 的数据训练集中,x_i 是指从文章的文本内容中提取的文本特征在跨媒体相关性学习得到的文本最大相关性子空间中的向量表示. y_i 是概念标签,其值也是由公式(10)确定的主题概率分布中每个主题的排序结果. 需要注意的是,对于每个概念 c,标签 y_i 的值在训练图像语义映射和文本语义映射的过程中是相同的,从而在图像语义映射和文本语义映射之间产生我们所期望的一致性匹配. 文本概念分类器 $w_c\in\mathbb{R}^{D_T}$ 在训练集 D_c 上通过最小化公式(11)所示的目标函数来进行训练的.

4　实验与结果

我们使用本文提出的方法构建了一个跨媒体语义检索模型,并在数据集上测试了该方法.

4.1　数据集

实验用来评估跨媒体相关性学习性能主要使用维基百科文本-图像数据集[1,2].

维基百科文本图像数据集是以维基百科的“特色文章”生成的,其中总共包含分为 10 个不同语义类别的 2 866 个图像文本对. 在每个文本图像对中,文本是描述人、地点或一些事件的文章,并且图像是与文章的内容密切相关的,每对数据都被标记为 10 个语义类别之一. 随机将数据集分为两个子数据

集:2 173 对为训练数据,693 对为测试数据. 对图像提取 SIFT 特征后,用 k-means 聚类方法来获取图像词袋特征,词频被用来得到文本单词的词袋特征[1,2].本文中每幅图像采用 1 000 维的图像视觉词袋,每个文本用 3 000 维的词袋特征进行表示.

4.2 度量标准

为了评估跨媒体相关性学习与检索方法的性能,进行两个跨媒体检索任务:图像检索文本以及文本检索图像. 本文采用平均精度均值(mean average precision, MAP)测试算法的整体性能. 如果返回结果中的数据和查询数据属于同一个类别标签则为相关,否则为不相关. 给定一个查询(图片或者文本)和返回 R 个检索结果,则精度均值为

$$\mathrm{AP}=\frac{1}{T}\sum_{r=1}^{R}P(r)\delta(r) \tag{12}$$

式中,T 为检索结果中与查询相关数据的个数,$P(r)$ 表示返回的前 r 个数据的准确率(即相关的数据所占的百分数). 如果第 r 个数据和查询数据相关,则 $\delta(r)=1$, 否则 $\delta(r)=0$. 然后通过对查询集中的所有查询的 AP 值求平均值来计算 MAP 的值. 其中,我们将 R 的值设为 50,MAP 值越大,算法准确性越高.

4.3 实验结果

我们进行 3 组实验来验证本文方法的有效性. 第 1 组实验验证改进结构性能上的有效性;第 2 组实验验证 CMC-DCCA 在跨媒体相关性学习上的有效性;第 3 组实验将 CMSCR 与目前主流的跨媒体检索方法进行对比.

4.3.1 不同激活函数的实验

采用不同激活函数的结果模型在维基百科文本-图像数据集的性能比较如图 5 所示。

实验中采用 3 层网络,其中 sigmoid 函数、tanh 函数、ReLu 函数实验运用公式(7)时,式中的 α 值设为 0;线性函数以及线性函数结合立方根 sigmoid 函数实验中,α 的值设为 1. 其中立方根 sigmoid 函数没用文[13]中采用的非饱和非线性函数. 从图 5 可以看出,采用线性函数结合立方根 sigmoid 函数在维基百科数据集上的性能显著优于其他激活函数.

实验中对改进的网络结构与传统结构训练过程的迭代时间进行比较. 在维基百科数据集上,训练阶段达到 L-BFGS 算法收敛,改进结构的平均迭代次数为 193.2,传统结构的平均迭代次数为 226. 改进后的方法所需的平均迭代次数较少.这是由于线性投影损失层将原始输入数据进行初步的特征提取,使得上层的非线性隐含层获得了一个较好的特征输入,从而加快了收敛速度.因此,如果对原始的网络结构加入适当的线性变换,改进后的网络结构可以为深层网络优化的其他问题提供更快更好的解决方案.

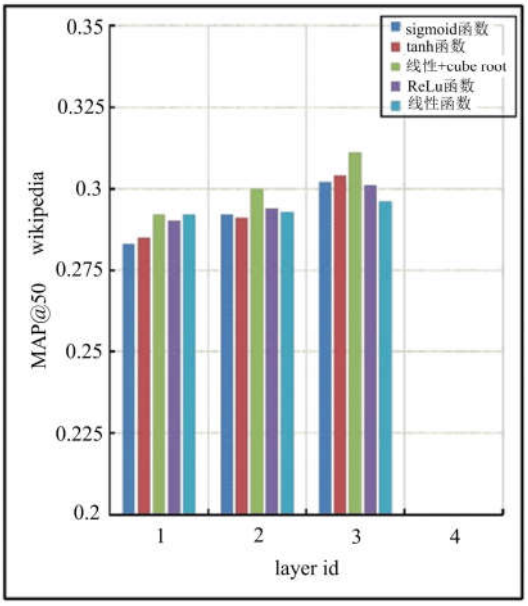

图 5 不同激活函数的结果对比

Fig.5 MAP@50 of different activation function

4.3.2 不同相关性学习方法的实验

我们将 CMC-DCCA 模型与以下相关性学习方法进行比较:典型相关性分析(CCA)[1],核化典型相关性分析(KCCA)[5],混合概率相关性分析(MixPCCA)[8],深度典型相关性分析(DCCA)[13].

各种相关性学习算法在维基百科数据集上的 MAP 结果如表 1 所示,标粗结果表示最佳结果. 在实验中,我们提交图像数据来检索文本或者提交文本数据来检索图像. 从中我们可以看出 CMC-DCCA 在所有的相关性学习方法中获得最佳的效果.

表 1 不同相关性学习方法的 MAP 结果

Tab.1 MAP@50 of different correlation learning methods

methods	Wikipedia		
	image query	text query	average
CCA	0.210	0.194	0.202
KCCA	0.287	0.279	0.283
DCCA	0.272	0.263	0.268
MixPCCA	0.295	0.313	0.304
CMC-DCCA	**0.318**	**0.339**	**0.329**

KCCA 的性能优于 CCA 说明在处理跨媒体相关性学习时，非线性模型要比线性模型效果好，所以构建非线性模型是能提高 CCA 性能的有效方法，因此，DCCA、MixPCCA 和 CMC-DCCA 均针对挖掘数据中存在的非线性关系对 CCA 进行了扩展. 其中 CMC-DCCA 效果优于所有这些基于 CCA 的方法，也印证了跨媒体数据中不存在简单的线性相关或者非线性相关关系.

4.3.3　与其他跨媒体检索方法的对比

我们将 CMSCR 模型与目前效果较好的跨媒体模型进行比较：CMC-DCCA，Corr-AE[11]，Corr-Cross-AE[11]，基于全模态对应自动编码机的跨模态检索算法(corr-full-AE)[11]以及 RE-DNN[12].

几种不同跨媒体检索方法在维基百科数据集的结果如表 2 所示，其中对性能前 3 的结果进行加粗表示. 从表中可以明显看出，与其他深度学习方法如 corr-cross-AE、corr-full-AE、RE-DNN 相比，跨媒体深度相关性分析 CMC-DCCA 的效果要略微低一点.

表 2　不同跨媒体检索方法的结果对比

Tab.2　The results of different cross-media retrieval methods

methods	wikipedia		
	image query	text query	average
CMC-DCCA	0.318	0.339	0.329
corr-AE	0.326	**0.361**	0.344
corr-cross-AE	**0.336**	0.341	0.338
corr-full-AE	0.335	**0.368**	**0.352**
RE-DNN	**0.341**	0.353	**0.347**
CMSCR	**0.348**	**0.359**	**0.354**

在跨媒体深度相关性学习中加入语义映射以后的 CMSCR 效果是所有方法中最优的. 从而证明了所提 CMSCR 方法在跨媒体检索应用上的有效性.

本文对采用 CMSCR 方法在维基百科数据集上进行跨媒体语义相关性检索的示例如图 6 所示. 图 6 分别展示了 CMSCR 用于图像检索文本以及文本检索图像的示例.

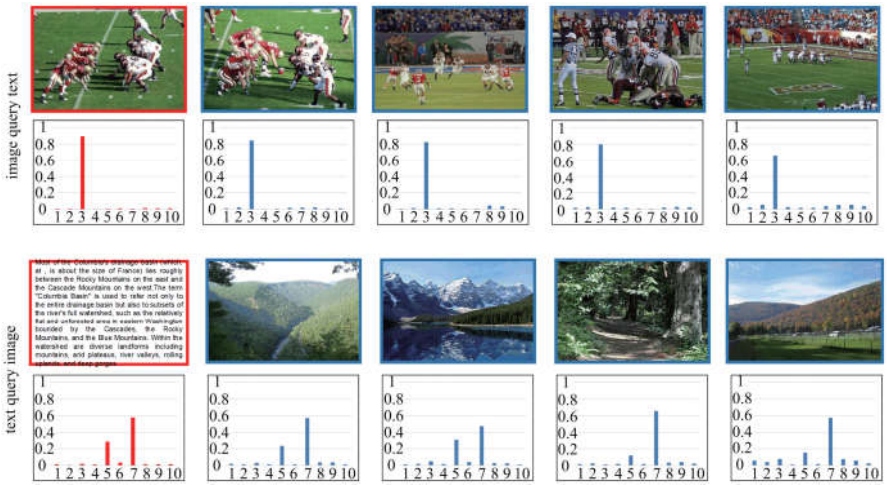

图 6　CMSCR 在维基百科数据集上的跨媒体检索示例

Fig.6　A cross-media retrieval result on Wikipedia dataset with CMSCR

图 6 上部为查询图像以及检索得到的文本所对应的图像，上部下面为查询图像的语义分布直方图以及检索得到的文本数据对应的不同语义分布的直方图.图 6 以“运动”类别中的橄榄球运动为例进行检索，返回的文本均为描述橄榄球比赛的文本，从图 6 下部的语义直方图中也能看出检索的结果是十分相关和准确的.

图 6 下部为查询文本以及检索得到的图像，下部下面为查询文本的语义分布直方图以及检索得到的图像数据对应的不同语义分布的直方图. 图中以“地理”类别中的一个描述国家地质及森林公园的文本为例进行检索. 语义分布图中，id 5 的语义对应“艺术”类别，id 7 的语义对应“地理”类别. 从检索得到的结果可以看出，前两个结果在语义分布上与查询文本属于同一类别并且相似程度很高. 后两个结果的语义分布中更偏向于“地理”类别，而在“艺术”类别上的语义较低. 从整体的结果来看，检索返回的结果和输入属于同一类别并且相相似程度较高.

5 结论

本文针对深度典型相关性分析进行结构改进，并用其构建跨媒体深度相关性学习模型. 其中线性投影层的训练是同非线性隐含层的训练结合在一起的，以确保得到相对较少并且更加抽象和更加准确的特征表示. 另外，我们提出一种语义对齐的数据驱动方法，根据从多媒体数据中发现的词汇概念及其标签信息分别对图像和文本数据训练语义映射来实现跨媒体检索. 在维基百科文本图像数据集的实验结果表明，本文提出的跨媒体语义检索算法(CMSCR)是有效的.

参考文献(References)

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

[2] PEREIRA J C, COVIELLO E, DOYLE G, et al. On the role of correlation and abstraction in cross-modal multimedia retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(3): 521-535.

[3] WANG S, LU J, GU X, et al. Unsupervised discriminant canonical correlation analysis based on spectral clustering[J]. Neurocomputing, 2016, 171(C): 425-433.

[4] ZU C, ZHANG D. Canonical sparse cross-view correlation analysis[J]. Neurocomputing, 2016, 191: 263-272.

[5] BALLAN L, URICCHIO T, SEIDENARI L, et al. A cross-media model for automatic image annotation [C]// Proceedings of International Conference on Multimedia Retrieval. New York: ACM Press, 2014: No.73(1-8).

[6] WANG S, ZHUANG F, JIANG S, et al. Cluster-sensitive structured correlation analysis for web cross-modal retrieval [J]. Neurocomputing, 2015, 168: 747-760.

[7] GONG Y, KE Q, ISARD M, et al. A multi-view embedding space for modeling internet images, tags, and their semantics [J]. International Journal of Computer Vision, 2014, 106(2): 210-233.

[8] 张博，郝杰，马刚，等. 混合概率典型相关性分析[J]. 计算机研究与发展，2015，52(7):1463-1476.
ZHANG B, HAO J, MA G, et al. Mixture of probabilistic canonical correlation analysis[J]. Journal of Computer Research and Development, 2015, 52(7): 1463-1476.

[9] 张博，郝杰，马刚，等. 基于弱匹配概率典型相关性分析的图像自动标注[J]. 软件学报，2017，28(2)：292-309.
ZHANG B, HAO J, MA G, et al. Automatic image annotation based on semi-paired probabilistic canonical correlation analysis [J]. Journal of Software, 2017, 28(2): 292-309.

[10] SRIVASTAVA N, SALAKHUTDINOV R. Learning representations for multimodal data with deep belief nets [C]// International Conference on Machine Learning Workshop. Edinburgh, Scotland: IMLS Press, 2012: 1-8.

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

[12] WANG C, YANG H, MEINEL C. Deep semantic mapping for cross-modal retrieval [C]// 27th International Conference on Tools with Artificial Intelligence. Vietri sul Mare, Italy: IEEE Computer Society, 2015: 234-241.

[13] ANDREW G, ARORA R, BILMES J A, et al. Deep canonical correlation analysis[C]// Proceedings of the 30th International Conference on Machine Learning . Atlanta, USA: IMLS Press, 2013: 1247-1255.

[14] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, (3): 993-1022.

[15] LIU D C, NOCEDAL J. On the limited memory BFGS method for large scale optimization[J]. Mathematical programming, 1989, 45(1): 503-528.

[16] JOACHIMS T. Optimizing search engines using clickthrough data[C]// Proceedings of the 8th ACM SIGKDD International Conference on Knowledge discovery and Data Mining. Edmonton, Canada: ACM Press, 2002: 133-142.