基于协同矩阵分解的单标签跨模态检索

李新卫,吴 飞,荆晓远
(南京邮电大学 自动化学院,江苏 南京 210023)

摘 要:针对跨模态检索存在的存储空间大、检索速度慢等缺点,提出了一种基于协同矩阵分解单标签跨模态检索方法,目标函数主要由协同矩阵分解、哈希函数和保持局部流形几何结构的图正则化三部分组成。矩阵分解学习训练数据集在低维潜在语义空间的哈希编码的简洁表示;哈希函数用来学习投影,将训练集外的样本表示成学习到的子空间的哈希码,根据汉明排序进行相似性搜索;图正则化用来保持原始空间的局部流行几何结构,该算法将这三部分有机地结合起来。为了证实该算法的有效性,在两个常用的数据集 Wiki 和 Pascal VOC 2007 进行了大量的实验,并与一些常用的相关方法进行了比较,结果证明了该算法的优越性。

关键词:协同矩阵分解;哈希函数;图正则化;稀疏图;跨模态检索
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2018)11-0099-04
doi:10.3969/j.issn.1673-629X.2018.11.022

Cross-modality Retrieval Based on Collective Matrix Factorization with Single Label

LI Xin-wei, WU Fei, JING Xiao-yuan
(School of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: Aiming at the disadvantages of large storage space and slow retrieval speed, we propose a novel cross-modality retrieval algorithm based on collective matrix factorization with single label. The objective function is mainly composed of cooperative matrix factorization, hash function and graph regularization of local manifold geometry structure. In particular, matrix decomposition learns the concise representation of hash coding of training data sets in low-dimensional latent semantic space. The hash function is used to learn the projection, the samples outside the training set are represented as the hash code of the learned subspace, and the similarity search is conducted according to Hamming sequence. Graph regularization is used to maintain the local popular geometry of the original space. We conjecture all these to improve the retrieval accuracy. Experiment on two cross-modality visual search datasets, Wiki and Pascal VOC 2007, shows that the proposed algorithm can significantly outperform the various state-of-the-art relevant methods.

Key words: collective matrix factorization; hash function; graph regularization; sparse graph; cross-modality retrieval

0 引 言

随着互联网的迅速发展,社会步入了大数据时代。大数据通常以图像、文本等多种不同的模态表示,而模态间的数据并不是独立的,而是有着本质的联系,如何挖掘出数据之间的关联信息成为了人们关注的热点。跨模态检索[1-3]作为一种基本的相关技术,在机器学习、计算机视觉和数据挖掘等领域应用广泛,然而大数据具有数据量大、维度高以及模态间的语义鸿沟[4]等一系列特点,从而使得针对大数据的跨模态检索困难重重。

为了减轻模态间的差异性,相关学者提出了一系列方法,一部分关注于潜在子空间学习,主要是用来学习两个投影,将文本与图像数据投影到公共空间进行相关性分析,比如 CCA[5] 及其变形[6];而哈希算法[7-9]作为一种近似最近邻检索技术,具有存储量小、检索速度快等特点,典型方法有 CVH[10]、IMH[11] 和 SCM[12]。然而,这些方法都有局限性,比如检索精度低、速度慢,因此设计更好的算法是相关工作者亟需解决的难题。基于此,文中提出一种基于协同矩阵分解的单标签跨模态检索方法。

收稿日期:2017-12-02 **修回日期**:2018-04-05 **网络出版时间**:2018-05-28
基金项目:国家自然科学基金(61702280);江苏省自然科学基金(BK20170900);江苏省高等学校自然科学研究项目(17KJB520025);南京邮电大学引进人才科研启动基金(NY217009)
作者简介:李新卫(1991-),男,硕士研究生,研究方向为模式识别、跨模态检索;荆晓远,教授,博导,研究方向为模式识别、人工智能。

1 方法描述

基于协同矩阵分解的单标签跨模态哈希检索方法由三部分构成:矩阵分解、哈希函数和图正则化项。矩阵分解学习训练集在低维潜在语义子空间的哈希码表示;哈希函数学习投影,将训练集外的样本表示成低维的哈希码;图正则化是用来保持原数据的局部几何结构的稀疏图[9]。

1.1 协同矩阵分解

协同矩阵分解主要用于低秩表示学习[13]。假设图像模态为 X_1,文本模态为 X_2,学习基矩阵 $\boldsymbol{U}_1 \in R^{d_1\times r}$、$\boldsymbol{U}_2 \in R^{d_2\times r}$,将 X_1 和 X_2 投影到 r 维的二进制语义空间 $V \in \{0,1\}^{r\times N}$。其中 r 为二进制哈希的长度,V 表示不同模态的公共表示。一般来讲,该过程可以通过下列目标函数求得:

$$\begin{cases}\min\limits_{U_1,U_2,V}\sum\limits_{i=1}\lambda_i \parallel X_i - \boldsymbol{U}_i V\parallel_F^2\\ \text{s.t. } \boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}} = \boldsymbol{I},\ \forall\ i, V\in\{0,1\}^{r\times N}\end{cases} \tag{1}$$

其中,λ_i 表示该模态的权重系数,满足 $\sum\limits_{i=1}\lambda_i = 1$。

1.2 哈希函数

设训练集外的样本为 x,学习两个哈希函数分别将 x 的图像 $x_1 \in R^{d_1}$ 和文本 $x_2 \in R^{d_2}$ 进行二进制编码。不妨假设采用简单的线性函数

$$h(x_i) = \operatorname{sign}(x_i\boldsymbol{P}_i + \boldsymbol{b}_i) \tag{2}$$

其中,$\operatorname{sign}(x)$ 为符号函数;$\boldsymbol{P}_i \in R^{d_i\times r}$ 为映射矩阵;$\boldsymbol{b}_i$ 为用来保证哈希编码平衡的偏置向量。

通过哈希函数,原始的图像 x_1 和对应的文本 x_2 在低维的潜在语义空间分别表示为:

$$\begin{cases}V_1 = h(x_1)\\ V_2 = h(x_2)\end{cases} \tag{3}$$

根据以上假设,相似的样本经过编码后的哈希距离应尽可能小,因此具有目标函数:

$$\min_{P_1,P_2}\parallel V - V_1\parallel_F^2 + \parallel V - V_2\parallel_F^2 = \parallel V - h(X_1)\parallel_F^2 + \parallel V - h(X_2)\parallel_F^2 \tag{4}$$

1.3 多模态图正则化项

图正则项[14-15]在机器学习、计算机视觉等领域取得了不错的发展,其用来维护局部几何结构,保证模态内相似性和模态间相似性。

模态间相似性:不同模态具有不同的特征表示,但是同一样本共享相同的语义表示。为了在低维语义中能够保持模态间的相似性,定义图像和文本的模态间相似性矩阵:

$$\boldsymbol{W}_{12}^{ij} = \begin{cases}1, & X_1^i \text{ 和 } X_2^j \text{ 属于同一类}\\ 0, & \text{其他情况}\end{cases} \tag{5}$$

模态内相似性:单模态相似的实例投影到低维语义中应该保持近邻关系,即哈希码的关联性尽可能大。定义单模态 KNN 相似矩阵:

$$\boldsymbol{W}_k^{ij} = \begin{cases}\exp(-z_k^{ij}/2\sigma^2), & X_k^i \in N(X_k^j) \text{ 或者 } X_k^j \in N(X_k^i)\\ 0, & \text{其他情况}\end{cases} \tag{6}$$

其中,Z_k^{ij} 表示 X_k^i 和 X_k^j 的欧氏距离,即 $Z_k^{ij} = |X_k^i - X_k^j|^2$。

整体的相似性矩阵为:

$$\boldsymbol{W} = \begin{bmatrix}\beta\boldsymbol{W}_1 & \boldsymbol{W}_{12}\\ \boldsymbol{W}_{21} & \beta\boldsymbol{W}_2\end{bmatrix} \tag{7}$$

其中,β 为保证模态间相似性和模态内相似性平衡的参数;$\boldsymbol{W}_1$、$\boldsymbol{W}_2$ 分别为图像和文本的单模态相似性矩阵;$\boldsymbol{W}_{12}$、$\boldsymbol{W}_{21}$ 为模态间相似性矩阵,$\boldsymbol{W}_{12} = \boldsymbol{W}_{21}$。

因此,图正则项的目标函数可以表示为:$\sum\limits_{i=1}^{2}\sum\limits_{j=1}^{2}\operatorname{tr}(\boldsymbol{X}_i\boldsymbol{P}_i\boldsymbol{L}_{ij}\boldsymbol{P}_i^{\mathrm{T}}X_i^{\mathrm{T}})$。其中,$\boldsymbol{L}_{ij}$ 为对应的拉普拉斯矩阵。

1.4 整体目标函数

综上所述,整体目标函数为:

$$\begin{aligned}\boldsymbol{\Phi} = &\sum_{i=1}\lambda_i \parallel \boldsymbol{X}_i - \boldsymbol{U}_i V\parallel_F^2 +\\ &\alpha\sum_{i=1}^{2}\parallel \boldsymbol{V} - h(X_i)\parallel_F^2 +\\ &\gamma\sum_{i=1}^{2}\sum_{j=1}^{2}\operatorname{tr}(\boldsymbol{X}_i\boldsymbol{P}_i L_{ij}\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{X}_i^{\mathrm{T}})\\ \text{s.t. }&\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}} = \boldsymbol{I},\ \forall\ i\end{aligned} \tag{8}$$

其中,α 和 γ 分别为对应的权重因子。

2 算法求解步骤

目标函数 $\boldsymbol{\Phi}$ 整体来说是非凸的,是个 *NP* 难问题,然而对于其中一个参数是可解的,因此可以采用迭代优化算法求解,具体步骤如下:

步骤 1:更新 $\boldsymbol{U}_1$、$\boldsymbol{U}_2$,固定 V、$\boldsymbol{P}_1$ 和 $\boldsymbol{P}_2$,通过拉格朗日乘子法可得:

$$\begin{cases}\boldsymbol{U}_1 = \boldsymbol{X}_1V^{\mathrm{T}}\left(2\dfrac{\eta_1}{\lambda_1}\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{U}_1 - \dfrac{\eta_1}{\lambda_1}I - VV^{\mathrm{T}}\right)^{-1}\\ \boldsymbol{U}_2 = \boldsymbol{X}_2V^{\mathrm{T}}\left(2\dfrac{\eta_2}{\lambda_2}\boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{U}_2 - \dfrac{\eta_2}{\lambda_2}I - VV^{\mathrm{T}}\right)^{-1}\end{cases} \tag{9}$$

步骤 2:更新 V,固定 $\boldsymbol{U}_1$、$\boldsymbol{U}_2$、$\boldsymbol{P}_1$ 和 $\boldsymbol{P}_2$,目标函数可以写成:

$$\begin{aligned}\boldsymbol{\Phi} = &\lambda_1\parallel X_1 - \boldsymbol{U}_1V\parallel_F^2 + \lambda_2\parallel X_2 - \boldsymbol{U}_2V\parallel_F^2 +\\ &\alpha(\parallel V - h(X_1)\parallel_F^2 + \parallel V - h(X_2)\parallel_F^2)\\ \text{s.t. }&V\in\{0,1\}^{r\times N}\end{aligned} \tag{10}$$

由于 V 是离散约束的,直接求解很棘手,故对其进

行松弛变换,将 $V \in \{0,1\}^{r\times N}$ 松弛为 $0 \leqslant V \leqslant 1$,通过拉格朗日乘子法可得:

$$\frac{\partial \boldsymbol{\Phi}}{\partial V} = [\lambda_1 \boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{U}_1 + \lambda_2 \boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{U}_2 + 4\alpha \boldsymbol{I}]V - 2[\lambda_1 \boldsymbol{U}_1^{\mathrm{T}}X_1 + \lambda_2 \boldsymbol{U}_2^{\mathrm{T}}X_2 + \alpha h(X_1) + \alpha h(X_2)]$$
$$\text{s.t. } \boldsymbol{\Psi}_{ij}V_{ij} = 0,\ V \geqslant 0 \tag{11}$$

利用 KKT 条件,得到 V 的更新公式为:

$$V_{ij} = V_{ij}\frac{2[\lambda_1 \boldsymbol{U}]_1^{\mathrm{T}}X_1 + \lambda_2 \boldsymbol{U}_2^{\mathrm{T}}X_2 + \alpha h(X_1) + \alpha h(X_2)]_{ij}}{[\lambda_1 \boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{U}_1 V + \lambda_2 \boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{U}_2 V + 4\alpha V]_{ij}} \tag{12}$$

步骤 3:更新 $\boldsymbol{P}_1$、$\boldsymbol{P}_2$,由拉格朗日乘子法解得:

$$\begin{cases}\boldsymbol{P}_1 = (2VX_1^{\mathrm{T}} - \boldsymbol{P}_2X_2L_{12}X_1^{\mathrm{T}})(2X_1X_1^{\mathrm{T}} + 4\gamma X_1L_{11}X_1^{\mathrm{T}})^{-1}\\ \boldsymbol{P}_2 = (2VX_2^{\mathrm{T}} - \boldsymbol{P}_1X_1L_{21}X_2^{\mathrm{T}})(2X_2X_2^{\mathrm{T}} + 4\gamma X_2L_{22}X_2^{\mathrm{T}})^{-1}\end{cases} \tag{13}$$

3　实验及结果分析

3.1　数据库及评价指标

为了验证该方法的有效性,在 Wiki 和 Pascal VOC 2007 数据集上与若干相关方法进行了对比,包括 CCA[4]、IMH[11]、CVH[10]、SCM_orth 和 SCM_seq[12]。

Wiki 数据集[16]包含 2 866 多媒体数据,分为 10 个主题,比如战争、艺术、天空等,其中每个样本是图像-文本对,图像是由 128 维的 BOVW SIFT 特征表示,文本是由 10 维的主题向量构成。

Pascal VOC 2007 数据集[17]包含 5 011/4 952 图像文本对,分为 20 类。部分图像是多标签的,文中只研究单标签,因此对该数据集进行相应处理,图像由 512 维的 Gist 特征表示,文本对应 319 维的词频特征。

实验中进行了两种跨模态检索任务:以图检文,Img2Text,即用图像去检索相关的文本;以文检图,Text2Img,即用文本去检索对应的图像。为了评估检索精度,使用 mAP[18]作为性能指标,其公式为:

$$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{AP}(q_i) \tag{14}$$

其中,q_i 为一查询样本;N 为查询样本量;AP( ) 的计算公式为:

$$\mathrm{AP}(q) = \frac{1}{T}\sum_{r=1}^{R}P_q(r)\xi(r) \tag{15}$$

其中,T 表示检索集中与 q 相关的总量;R 表示检索到的样本量;$\xi(r)$ 为指示函数。

3.2　实验结果

实验 1:在 Wiki 数据集上进行了 Img2Text 和 Text2Img 实验,随机抽取 2 173 个样本对作为训练集,其余的 695 个样本为测试集,实验结果如表 1 所示。

表 1　Wiki 数据集下各方法的 mAP

检索类型	对比方法	哈希码长度			
		16 位	32 位	64 位	128 位
Img2-Text	CCA	0.171 4	0.154 7	0.146 5	0.128 6
	IMH	0.195 2	0.200 3	0.208 4	0.209 7
	CVH	0.171 8	0.157 3	0.150 8	0.135 1
	SCM_orth	0.155 1	0.140 6	0.137 5	0.124 7
	SCM_seq	0.234 1	0.241 0	0.245 5	0.256 3
	文中方法	0.184 6	0.187 4	0.247 1	0.265 9
Text2-Img	CCA	0.159 0	0.141 5	0.130 0	0.115 1
	IMH	0.150 8	0.158 1	0.163 6	0.166 8
	CVH	0.150 5	0.134 7	0.121 4	0.113 6
	SCM_orth	0.155 4	0.139 8	0.129 3	0.113 7
	SCM_seq	0.225 7	0.245 9	0.243 7	0.247 6
	文中方法	0.201 2	0.542 8	0.624 7	0.676 3

实验 2:在 Pascal VOC 2007 数据集上进行了 Img2Text 和 Text2Img 实验,训练集为 2 808 对样本,测试集为 2 841 对样本,实验结果如表 2 所示。

表 2　Pascal VOC 2007 数据集下各方法的 mAP

检索类型	对比方法	哈希码长度			
		16 位	32 位	64 位	128 位
Img2-Text	CCA	0.172 8	0.165 2	0.188 6	0.211 4
	IMH	0.196 2	0.206 5	0.206 9	0.208 6
	CVH	0.173 2	0.170 5	0.187 7	0.203 5
	SCM_orth	0.160 3	0.169 5	0.190 2	0.235 8
	SCM_seq	0.168 2	0.172 8	0.220 6	0.254 1
	文中方法	0.162 1	0.182 4	0.238 7	0.265 4
Text2-Img	CCA	0.223 1	0.270 2	0.292 0	0.335 6
	IMH	0.231 6	0.287 8	0.310 9	0.345 8
	CVH	0.240 7	0.286 3	0.313 5	0.347 0
	SCM_orth	0.245 7	0.289 6	0.326 2	0.358 1
	SCM_seq	0.253 9	0.303 7	0.339 7	0.354 2
	文中方法	0.260 3	0.535 2	0.626 9	0.679 2

由表 1、表 2 可知,文中方法比其他方法效果好;随着哈希码变长,检索精度也越来越好,说明了长的哈希码更能够保持结构特征。

4　结束语

提出了一种基于协同矩阵分解的单标签跨模态哈希检索方法,目标函数由三部分构成,即矩阵分解、哈希函数和图正则化项。矩阵分解学习训练集在低维潜在语义子空间的哈希码;哈希函数用来学习投影,将训练集外的样本表示成低维空间上的哈希码,进行相似性搜索;图正则化用来保持原数据的局部流行几何结

构。在两种常用的数据集上进行了大量的实验,结果证明了该方法的优越性。

参考文献:

[1] 王开业. 跨模态数据分析与应用研究[D]. 北京:中国科学院大学,2015.

[2] 邓正恒. 跨模态信息检索方法的研究与实现[D]. 上海:复旦大学,2013.

[3] 丁 恒,陆 伟. 基于相关性的跨模态信息检索研究[J]. 现代图书情报技术,2016,32(1):17-23.

[4] 张 菁,沈兰荪,David Dagan Feng. 基于视觉感知的图像检索的研究[J]. 电子学报,2008,36(3):494-499.

[5] HARDOON D R,SZEDMAK S,SHAWE-TAYLOR J. Canonical correlation analysis:an overview with application to learning methods[J]. Neural Computation,2004,16(12):2639-2664.

[6] GONG Yunchao,KE Qifa,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.

[7] 叶卫国,韩水华. 基于内容的图像 Hash 算法及其性能评估[J]. 东南大学学报:自然科学版,2007,37:109-113.

[8] 赵玉鑫,刘光杰,戴跃伟,等. 一种新的视觉 Hash 算法[J]. 光学精密工程,2008,16(3):551-557.

[9] 金仲明. 基于哈希算法的海量多媒体数据检索研究[D]. 杭州:浙江大学,2015.

[10] KUMAR S,UDUPA R. Learning hash functions for cross-view similarity search[C]//Proceedings of the twenty-second international joint conference on artificial intelligence. Barcelona,Catalonia,Spain:AAAI Press,2011:1360-1365.

[11] SONG Jingkuan,YANG Yang,YANG Yi,et al. Inter-media hashing for large-scale retrieval from heterogeneous data sources[C]//Proceedings of the 2013 ACM SIGMOD international conference on management of data. New York:ACM,2013:785-796.

[12] ZHANG Dongqing,LI Wujun. Large-scale supervised multimodal hashing with semantic correlation maximization[C]//Twenty-eighth AAAI conference on artificial intelligence. Québec City,Québec,Canada:AAAI Press,2014:2177-2183.

[13] 陈 芸,吴 飞,荆晓远. 鲁棒低秩稀疏表示的在线目标跟踪[J]. 计算机工程与设计,2016,37(4):1062-1066.

[14] 张志武,荆晓远,吴 飞. 基于非负稀疏图的协同训练软件缺陷预测[J]. 计算机技术与发展,2017,27(7):38-42.

[15] WANG Kaiye,HE Ran,WANG Liang,et al. Joint feature selection and subspace learning for cross-modal retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2016,38(10):2010-2023.

[16] 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,2010:251-260.

[17] 李 萌. 基于特征选择的 Fisher 向量在图像分类中的应用[D]. 北京:北京交通大学,2014.

[18] LIU Hong,JI Rongrong,WU Yongjian,et al. Supervised matrix factorization for cross-modality hashing[C]//Proceedings of the twenty-fifth international joint conference on artificial intelligence. New York,NY,USA:AAAI Press,2016:1767-1773.

(上接第 98 页)

l.]:[s. n.],2005.

[8] 程 静. 基本情感生理信号的非线性特征提取研究[D]. 重庆:西南大学,2015.

[9] 谢 勇,徐健学,杨红军,等. 皮层脑电时间序列的相空间重构及非线性特征量的提取[J]. 物理学报,2002,51(2):205-214.

[10] 孙长城,王春方,王勇军,等. 脑卒中后抑郁症静息脑电信号非线性特征提取与分析[J]. 国际生物医学工程杂志,2013,36(3):143-146.

[11] 柴京京. 运动诱发局部肌肉疲劳肌音信号非线性特性分析[D]. 西安:陕西师范大学,2009.

[12] ECKMANN J P,KAMPHORST S O,RUELLE D. Recurrence plots of dynamical systems[J]. Europhysics Letters,1987,4(9):973-977.

[13] 冯思源. 基于交叉和排序递归图的癫痫脑电 RQA 分析[D]. 南京:南京邮电大学,2015.

[14] ZBILUT J P,JR C L W. Recurrence quantification analysis[M]//Wiley encyclopedia of biomedical engineering.[s. l.]:John Wiley & Sons,Inc.,2006.

[15] WAGNER J,KIM N J,ANDRE E. From physiological signals to emotions:implementing and comparing selected methods for feature extraction and classification[C]//IEEE international conference on multimedia and expo. Amsterdam,Netherlands:IEEE,2005:940-943.

[16] 温万惠,刘光远,熊 勰. 基于生理信号的二分类情感识别系统特征选择模型和泛化性能分析[J]. 计算机科学,2011,38(5):220-223.