计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200600003

用于多模态语义分析的嵌入共识自动编码器

孙圣姿　郭炳晖　杨小博
北京航空航天大学大数据与脑机智能高精尖中心　北京 100191
鹏程实验室　广东 深圳 518055
教育部数学信息与行为重点实验室和北京航空航天大学数学科学学院　北京 100191
(znlx367@163.com)

摘　要　跨模态检索技术是一项近年来的研究热点。多模态数据具有异质性,而不同形式的信息之间又有着相似性。传统的单模态方法只能以一种方式重构原始数据,并未考虑到不同数据之间的语义相似性,不能进行有效的检索。因此,文中建立了一个跨模态嵌入共识自动编码器(Cross-Modal Semantic Autoencoder with Embedding Consensus,ECA-CMSA),将原始数据映射到低维共识空间以保留语义信息,学习出对应的语义代码向量,并引入参数来实现去噪。然后,考虑到各模态之间的相似性,采用自动编码器将特征投影关联到语义代码向量。此外,对低维矩阵进行正则化稀疏约束,以平衡重构误差。在4个多模态数据集上验证所提方法的性能,实验结果证明其查询结果有所提升,实现了有效的跨模态检索。进一步,ECA-CMSA还可以应用于与计算机和网络有关的领域,如深度学习和子空间学习。该模型突破了传统方法中的障碍,创新地使用深度学习方法将多模态数据转换为抽象的表达,使其可以获得更好的准确度和识别结果。

关键词:多模态检索;嵌入共识;自动编码器;稀疏正则
中图法分类号　TP39

Embedding Consensus Autoencoder for Cross-modal Semantic Analysis

SUN Sheng-zi,GUO Bing-hui and YANG Xiao-bo
Beijing Advanced Innovation Center for Big Data and Brain Computing,Beihang University,Beijing 100191,China
Peng Cheng Laboratory,Shenzhen,Guangdong 518055,China
LMIB and School of Mathematical Sciences,Beihang University,Beijing 100191,China

Abstract　Cross-modal retrieval has become a topic of popularity,since multi-data is heterogeneous and the similarities between different forms of information are worthy of attention. Traditional single-modal methods reconstruct the original information and lacks of considering the semantic similarity between different data. In this work,an Embedding Consensus Autoencoder for Cross-Modal Semantic Analysis is proposed,which maps the original data to a low-dimensional shared space to retain semantic information. Considering the similarity between the modalities,an automatic encoder is utilized to associate the feature projection to the semantic code vector. In addition,regularization and sparse constraints are applied to low-dimensional matrices to balance reconstruction errors. The high dimentional data is transformed into semantic code vector. Different models are constrained by parameters to achieve denoising. The experiments on four multi-modal data sets show that the query results are improved and effective cross-modal retrieval is achieved. Further,ECA-CMSA can also be applied to fields related to computer and network such as deep and subspace learning. The model breaks through the obstacles in traditional methods,and uses deep learning methods innovatively to convert multi modal data into abstract expression,which can get better accuracy and achieve better results in recognition.

Keywords　Cross-modal retrieval,Embedding consensus,Autoencoder,Sparse regularization

1　研究背景及意义

随着互联网技术的发展,大量多媒体数据不断涌现,给信息检索带来了巨大的挑战。数据来源形式包括文本、图像、视频和音频等[1],其中图像与文字的信息表达颇为常见。近年来,用户通常希望以适合自己的方式来获得需求的数据。传统的单模态检索无法解决二者的兼容问题,因为它们只能以相同的方式返回原始数据进行查询,不能满足检索需求。鉴

到稿日期:2020-05-30　返修日期:2020-09-09　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:科技创新2030-"新一代人工智能"重大项目(2018AAA0102301);国家自然科学基金(11671025);民机项目(MJ-F-2012-04)
This work was supported by the Technological innovation 2030-Artificial Intelligence Project(2018AAA0102301), National Natural Science Foundation of China(11671025) and Fundamental Research of Civil Aircraft(MJ-F-2012-04).
通信作者:郭炳晖(guobinghui@buaa.edu.cn)

于此，跨模态检索方法被提出并迅速发展，其目的是检索不同的信息模式，如检索部分带有文本的图像。因此，跨模态信息检索成为了一类热点，是解决上述问题的有效方法[2]。

跨模态检索可以利用各类数据查询不同形式的信息，执行跨模态检索的关键问题是需要考虑到不同形式的数据之间的语义相似性。不同模态之间的异质性使得该问题面临挑战。目前，已有的图像-文本跨模态检索方法包括成对模型、排序、映射和图嵌入[3]。概率模型、度量学习方法和子空间学习方法被应用在许多数据集上。概率方法通过对联合多模态数据分布进行建模，来学习多模态的相关性[4]；度量学习方法则学习计算出不同模态之间的距离度量[5]；一些经典的方法[2]将数据投影到一个公共空间。语义信息是查询时保留下来的重要信息，不同形式的数据具有不同的特征空间，但它们拥有相同的语义空间，具有相同语义的数据在各种模态下的形式都具有关联性。语义信息不仅可以用于表示多模态之间的关联程度，而且可以表示各个模态内部的联系。为了获得良好的检索结果，可以利用嵌入方法同时保留语义和原始特征信息。Zhou 等[6]提出潜在的语义稀疏散列方法(Latent Semantic Sparse Hashing，LSSH)，该方法结合稀疏编码和矩阵分解来获得潜在的共享语义空间。在深度方法中，通常用卷积神经网络(CNN)来生成图像，而语义部分则将每个单词进行特征嵌入，并通过文字 CNN 或递归神经网络(RNN)来生成文本[7]。

语义信息的缺乏导致检索结果有限。部分回归方法，如 LCFS[7] 和 LGCFL[8]，专注于保留语义信息。然而，以上两种方法只能用于处理单模态情形，而忽略了多模态信息中各标签之间的相关性。此外，它们将公共空间固定为标签空间，在数据规模较大时效率较低。

跨模态检索技术涉及与数学、统计学等相关的一些基础知识。为了更好地满足人们的检索需求，通过各种多媒体数据进行有效的信息查询，将其应用到深度学习、子空间学习等与计算机和网络相关的领域，这是一项具有重要应用价值的课题。

目前，国内外学者都对此进行了研究，跨模态的相似性学习引起了学术界的极大关注。但是，数据的异质性和语义差别的存在，使得这一项研究非常具有挑战性。目前，最为常见的两种度量方法分别为最大化相关性和最小化欧氏距离[9]。最大化相关性的典型方法是 CCA[10] 及其改进方法，学习出一个潜在空间，该空间将两种模态的投影特征之间的相关性最大化。文献[11]利用 CCA 得到了与人相对应的 2D 和 3D 面部图像的共享潜在空间。最小化欧氏距离的方法包括 PLS 和 BLM。Sharma 等[12]利用 PLS 实现了在不同姿势、高分辨率和低分辨率面部图像以及照片与草图之间的异质面部识别。双线性模型(BLM)被用于跨媒体检索和异构人脸识别[2]。

自动编码器是一种无监督的神经网络模型，它可以学习输入数据的隐含特征，该过程称为编码；同时用学习到的新特征重构出原始输入数据，该过程称为解码。自动编码器[13]是经过训练的模型，用于学习一组数据的潜在表示形式，它利用训练数据集，可以将输入的信息复制到输出，因此潜在的表示形式为有效属性。部分学者提出了关于自动编码器的变形方法。Lange 等[14]将深度自动编码器的训练(用于学习紧凑的特征空间)与 RL 算法(用于学习策略)相结合。Sainat 等[15]利用训练集对 AE-BN 模式进行应用。传统的自动编码器只是简单地寻求潜在表示形式以重建原始数据，而本文方法考虑了其与语义代码向量的相似性。

为了解决上述问题，以达到更为高效的信息检索效果，本文提出了一种学习方法，称为跨模态嵌入共识自动编码器(ECA-CMSA)。

(1)将成对的图像和文字数据嵌入映射到一个统一的空间中，称为嵌入共识，同时保留原始特征信息和语义信息。进一步地，通过特征提取，将数据转换为相应的语义代码向量。该方法压缩了高维数据的多标签空间，并且利用参数来实现去噪，可以去除冗余的信息。

(2)采用成对的编码器-解码器来学习特征投影，一组用于图像形式，一组用于文本形式。考虑到投影后信息之间的相似性，并将其关联到语义代码向量。进一步地，最小化目标函数，对矩阵进行正则化稀疏约束来平衡重构误差。

2 跨模态嵌入共识自动编码器

结合相关工作，本文对已有方法进行了改进，构造了一组跨模态嵌入共识自动编码器(ECA-CMSA)。首先将成对的图像-文字数据统一映射到一个低维嵌入空间，保留其流形结构，将原始信息转换为对应的语义代码向量。同时，不断更新共识矩阵和语义代码矩阵。然后，通过对图像和文本投影矩阵的学习，利用编码器将其分别关联到相应的语义代码上，再将解码器重新投影回高维数据。最后，对解码器进行正则化稀疏约束，引入平衡参数重构原始特征，对多模态信息进行较为有效的检索。该方法的具体流程如图 1 所示。

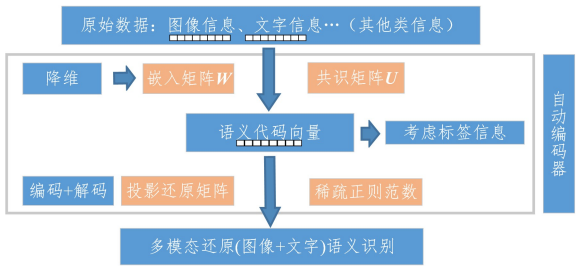


图 1 ECA-CMSA 方法的流程

Fig. 1 Process of ECA-CMSA

2.1 嵌入共识

令($\boldsymbol{V}$ $\boldsymbol{T}$)为原始的数据矩阵，其中，$\boldsymbol{V}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_N)^{\mathrm{T}}$ 为图像信息，$\boldsymbol{T}=(\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_N)^{\mathrm{T}}$ 为文字信息。向量($\boldsymbol{v}_i,\boldsymbol{t}_i$)代表第 i 行信息，(v_i^d,t_i^d)表示数据($\boldsymbol{v}_i,\boldsymbol{t}_i$)的第 d 个维度。考虑映射 $\varphi^d:(\boldsymbol{v}_i,\boldsymbol{t}_i)\rightarrow\boldsymbol{U}_i$，$\varphi^d\in D(i=1,2,\cdots,N)$，$\boldsymbol{U}_i$ 是映射后的低维共识向量，φ^d 为嵌入映射。嵌入降维后的共识矩阵 $\boldsymbol{U}_i$，实现了每一对图像与文字信息映射结果的统一，可以进一步学习出语义代码向量。流形降维可保留原始数据点的局部几何结构。为了防止结果受到噪声数据的影响，引入参数 γ_i^d，从而得到：

$$\begin{aligned}\Gamma_{C(i)}&=\sum_{d=1}^{D}\gamma_i^d\parallel\varphi^d(\boldsymbol{v}_i^d,\boldsymbol{t}_i^d)-\boldsymbol{U}_i\parallel^2\\&=\mathrm{diag}(\gamma_i)\,\mathrm{tr}\left(\begin{pmatrix}\varphi^1(\boldsymbol{v}_i^1,\boldsymbol{t}_i^1)-\boldsymbol{U}_i\\\cdots\\\varphi^D(\boldsymbol{v}_i^D,\boldsymbol{t}_i^D)-\boldsymbol{U}_i\end{pmatrix}(\varphi^1(\boldsymbol{v}_i^1,\boldsymbol{t}_i^1)-\boldsymbol{U}_i\cdots\right.\\&\quad\left.\varphi^D(\boldsymbol{v}_i^D,\boldsymbol{t}_i^D)-\boldsymbol{U}_i)\right)\\&=\mathrm{tr}\left(W_i\begin{pmatrix}-\mathrm{e}_{D+1}^{\mathrm{T}}\\\boldsymbol{I}_{D+1}\end{pmatrix}\mathrm{diag}(\gamma_i)(e_{D+1}\boldsymbol{I}_{D+1})W_i^{\mathrm{T}}\right)\\&=\mathrm{tr}(\boldsymbol{W}_i\boldsymbol{C}_i\boldsymbol{W}_i^{\mathrm{T}})\end{aligned}$$

其中,$\boldsymbol{W}_i=(\boldsymbol{U}_i\ \varphi^1(\boldsymbol{v}_i^1,\boldsymbol{t}_i^1)\cdots\ \varphi^D(\boldsymbol{v}_i^D,\boldsymbol{t}_i^D))$为低维嵌入矩阵,保留了原始信息的流形结构。$\varphi=\mathrm{diag}(\varphi^1\ \cdots\ \varphi^D)$,$\boldsymbol{C}_i=\begin{pmatrix}-e_{D+1}^{\mathrm{T}}\\I_{D+1}\end{pmatrix}diag(\gamma_i)(e_{D+1}\boldsymbol{I}_{D+1})$,将数据$(\boldsymbol{v}_i,\boldsymbol{t}_i)$通过嵌入共识矩阵转换为相应的语义代码向量。为了消除噪声的影响,当第$(\boldsymbol{v}_i,\boldsymbol{t}_i)$组的数据映射结果异常时,$\gamma_i^d$趋于 0。利用原始的图像和文字信息提取出对应的特征,将$\boldsymbol{W}_i$可以写成$\boldsymbol{W}_i=WE_i$的形式,其中特征矩阵为:$\boldsymbol{E}_i=(e_i^{\mathrm{T}}\ \cdots\ e_{N+(i-1)D+1}^{\mathrm{T}}\cdots\ e_{N+iD}^{\mathrm{T}})$

在第i组数据中,将表示图像和文字的N组成对的原始数据信息求和。

$$\begin{aligned}\Gamma_C&=\sum_{i=1}^{N}\Gamma_{C(i)}=\sum_{i=1}^{N}\mathrm{tr}(\boldsymbol{W}_i\boldsymbol{C}_i\boldsymbol{W}_i^{\mathrm{T}})\\&=\sum_{i=1}^{N}\mathrm{tr}(\boldsymbol{W}E_i\boldsymbol{C}_iE_i^{\mathrm{T}}\boldsymbol{W})=\mathrm{tr}(\boldsymbol{W}\boldsymbol{C}\boldsymbol{W}^{\mathrm{T}})\end{aligned}$$

其中,

$$\begin{aligned}\boldsymbol{W}&=(\boldsymbol{U}\varphi^1(\boldsymbol{v}_1^1,\boldsymbol{t}_1^1)\ \cdots\ \varphi^D(\boldsymbol{v}_1^D,\boldsymbol{t}_1^D)\ \cdots\ \varphi^1(\boldsymbol{v}_N^1,\boldsymbol{t}_N^1)\ \cdots\\&\quad\varphi^D(\boldsymbol{v}_N^D,\boldsymbol{t}_N^D)\end{aligned}$$

$$\boldsymbol{C}=\sum_{i=1}^{N}E_i\boldsymbol{C}_iE_i^{\mathrm{T}}=\boldsymbol{D}-\boldsymbol{\gamma}$$

其中,$\boldsymbol{\gamma}$是映射点与原始数据点的相关矩阵,$\boldsymbol{D}$为对角矩阵。利用矩阵$\boldsymbol{C}$,可以将图像和文字信息转换成对应的语义代码。进一步地,令$\Phi=(\varphi^1(\boldsymbol{v}_1^1,\boldsymbol{t}_1^1)\ \cdots\ \varphi^D(\boldsymbol{v}_1^D,\boldsymbol{t}_1^D)\ \cdots\ \varphi^1(\boldsymbol{v}_N^1,\boldsymbol{t}_N^1)\ \cdots\ \varphi^D(\boldsymbol{v}_N^D,\boldsymbol{t}_N^D))$,最终的表达式为:

$$\begin{aligned}\Gamma_C&=\mathrm{tr}(\boldsymbol{W}\boldsymbol{C}\boldsymbol{W}^{\mathrm{T}})\\&=\mathrm{tr}\left((U\ \ \Phi)\begin{pmatrix}\boldsymbol{C}_{11}&\boldsymbol{C}_{12}\\\boldsymbol{C}_{21}&\boldsymbol{C}_{22}\end{pmatrix}\begin{pmatrix}U^{\mathrm{T}}\\\Phi^{\mathrm{T}}\end{pmatrix}\right)\\&=\mathrm{tr}\boldsymbol{U}\boldsymbol{C}_{11}\boldsymbol{U}^{\mathrm{T}})+\mathrm{tr}(\Phi\boldsymbol{C}_{21}\boldsymbol{U}^{\mathrm{T}})+\mathrm{tr}(\boldsymbol{U}\boldsymbol{C}_{12}\Phi^{\mathrm{T}})+\mathrm{tr}(\Phi\boldsymbol{C}_{22}\\&\quad\Phi^{\mathrm{T}})\end{aligned}$$

2.2 嵌入共识

2.2.1 目标函数

在将图像和文本映射到嵌入式共识空间后,$\boldsymbol{U}_i$可以包含足够的原始数据信息。下面分别对投影矩阵$\boldsymbol{P}_v$,$\boldsymbol{P}_t$进行学习:利用编码器将图像和文本投影关联到语义代码向量$\boldsymbol{C}$上,对解码器进行约束,使得代码向量能够重构图像和文字的原始特征。编码器和解码器是线性对称的,两个编码器$\boldsymbol{P}_v$,$\boldsymbol{P}_t$将图像和文本投影到低维空间$\boldsymbol{A}$,两个解码器$\boldsymbol{P}_v^{\mathrm{T}}$,$\boldsymbol{P}_t^{\mathrm{T}}$将$\boldsymbol{A}$投影返回高维数据,隐藏层同时包含原始的图像和文本信息。

对于图像数据,采用自动编码器的嵌入形式表示原始特征的信息。图像-文本的成对表示形式需要统一,因为在检索阶段,当查询信息给定后,将根据相似程度进行排序查询。由此得到:

$$\boldsymbol{P}_v\boldsymbol{V}=\boldsymbol{A},\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{V}=\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{V}\boldsymbol{A}$$

其中,$\boldsymbol{A}\in\mathbb{R}^{d\times n}$将$N$组训练文本表示在$d$维的隐藏空间中,这种形式是一种具有约束力[19]的线性自动编码器,并且只有一个隐藏层。编码器将输入信息投影到较低的隐藏层,解码器将其投影回原始特征空间。

对于文本数据,为了使低维特征具有恢复原始信息点的能力,令:

$$\boldsymbol{P}_t\boldsymbol{T}=\boldsymbol{A},\boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{T}=\boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{T}\boldsymbol{A}$$

由以上两个表达式可以得到一个多模态自动编码器。此外,还需要确保隐藏层包含足够的语义信息。因此,将数据的隐藏表示与学习得到的语义代码向量$\boldsymbol{C}$相关联,考虑不同模态之间的相似性,利用标签信息规范自动编码器的潜在表示形式。在此基础上,加入对低维矩阵$\boldsymbol{A}$的正则化稀疏约束,从而得到最终的目标函数:

$$\begin{aligned}&\min_{P_v,P_t,A}\parallel\boldsymbol{P}_v\boldsymbol{V}+\boldsymbol{P}_t\boldsymbol{T}-\boldsymbol{A}\parallel_F^2+\beta(\parallel\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{V}-\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{V}\boldsymbol{A}\parallel_F^2+\\&\parallel\boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{T}-\boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{T}\boldsymbol{A}\parallel_F^2)+\eta\parallel\boldsymbol{A}-\boldsymbol{C}\parallel_F^2+\parallel\boldsymbol{A}\parallel_1\\&\text{s.t. }\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{P}_v=\boldsymbol{I},\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{V}L_A\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}_v=\boldsymbol{I},\\&\qquad\boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{P}_t=\boldsymbol{I},\boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{T}L_A\boldsymbol{T}^{\mathrm{T}}\boldsymbol{P}_t=\boldsymbol{I},\\&\qquad A_{ii}=0\end{aligned}$$

其中,β为平衡两类数据信息的权重参数,η为决定语义代码向量相对重要性的参数。

2.2.2 算法求解过程

采用交替迭代更新的方法,对目标函数分别进行求解。

(1)固定$\boldsymbol{A}$,更新$\boldsymbol{P}_v$,$\boldsymbol{P}_t$

投影矩阵$\boldsymbol{P}_v$,$\boldsymbol{P}_t$的求解方法类似,令:

$$L_A=(\boldsymbol{I}-\boldsymbol{A})(\boldsymbol{I}-\boldsymbol{A})^{\mathrm{T}}$$

将表达式整理成关于P_v的单变量函数:

$$\begin{aligned}&\min_{P_v}\parallel\boldsymbol{P}_v\boldsymbol{V}-\boldsymbol{A}\parallel_F^2+\beta\mathrm{tr}(\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{V}L_A\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}_v)\\&\text{s.t. }\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{P}_v=\boldsymbol{I},\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{V}L_A\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}_v=\boldsymbol{I},\boldsymbol{A}_{ii}=0\end{aligned}\tag{1}$$

同理,$\boldsymbol{P}_t$的更新函数为:

$$\begin{aligned}&\min_{P_t}\parallel\boldsymbol{P}_t\boldsymbol{T}-\boldsymbol{A}\parallel_F^2+\beta\mathrm{tr}(\boldsymbol{P}_t{}^{\mathrm{T}}\boldsymbol{T}\boldsymbol{L}_A\boldsymbol{T}^{\mathrm{T}}\boldsymbol{P}_t)\\&\text{s.t. }\boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{P}_t=\boldsymbol{I},\boldsymbol{P}_t^{\mathrm{T}}T\boldsymbol{L}_AT^{\mathrm{T}}\boldsymbol{P}_t=\boldsymbol{I},\boldsymbol{A}_{ii}=0\end{aligned}\tag{2}$$

(2)固定$\boldsymbol{P}_v$,$\boldsymbol{P}_t$,更新$\boldsymbol{A}$

对变量$\boldsymbol{A}$求偏导,得到:

$$\begin{aligned}&(2\boldsymbol{A}-\boldsymbol{P}_v\boldsymbol{V}-\boldsymbol{P}_t\boldsymbol{T})+2\beta(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}_v\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{V}+\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}_v\boldsymbol{P}_v^{\mathrm{T}}\boldsymbol{V}\boldsymbol{A}+\\&\boldsymbol{T}^{\mathrm{T}}\boldsymbol{P}_t\boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{T}+\boldsymbol{T}^{\mathrm{T}}\boldsymbol{P}_t\boldsymbol{P}_t^{\mathrm{T}}\boldsymbol{T}\boldsymbol{A})+\eta(\boldsymbol{A}-\boldsymbol{C})+\frac{\boldsymbol{A}}{\parallel\boldsymbol{A}\parallel_1}=0\end{aligned}\tag{3}$$

类似 LASSO 问题的求解方法,可以更新矩阵$\boldsymbol{A}$。

综合以上两个过程,本文提出了基于嵌入式映射共识的跨模态自动编码器算法,如算法 1 所示。

算法 1 基于嵌入式映射共识的跨模态自动编码器算法

输入:原始数据矩阵 **V**,**T** 和代码向量 **C**

输出:投影矩阵 $\mathbf{P}_\mathrm{v}$,$\mathbf{P}_\mathrm{t}$

初始化:参数 β,η

利用矩阵 **C** 初始化矩阵 **A**;

迭代更新:

固定 **A**,根据式(1)、式(2)更新 $\mathbf{P}_\mathrm{v}$,$\mathbf{P}_\mathrm{t}$;

固定 $\mathbf{P}_\mathrm{v}$,$\mathbf{P}_\mathrm{t}$,根据式(3)更新 **A**;

直到收敛。

3 本文方法的应用

本节将通过实验来验证本文方法的性能。在4个多模态数据集上，将ECA-CMSA方法与其他已有的方法进行对比，来验证其有效性。同时，对各类指标值进行具体的结果分析。

3.1 数据集和对比方法

接下来，本文利用WIKI[16]，TVGraz[17]，NUS-WIDE[18]和MIRFLICKR[19]4个多模态数据集，将ECA-CMSA方法与CCA[10]，BLM[2]，LCFS[7]，LGCFL[8]，JFSSL[3]5种已有的方法进行比较。

(1)WIKI。其来源于“Wikipedia featured articles”，包含10个语义类别的2900个图像-文本对。本文利用2200个图文对进行训练，700个图文对进行测试。文本由10维的DDA特征表示，图像由1001维的CNN特征表示。

(2)TVGraz。其包含2594个图像-文本对。本文在实验中选择至少10个单词和2500个图像-文本对，其中2000对用于训练，500对用于测试。每个图像均由4000维的特征表示，每个文本均由具有8300个顶点的图像表示。

(3)NUS-WIDE。其是一个实际的图像数据集，原始信息包含269648张具有81个概念的图像。每个图像都与用户标签相关联，这些用户标签可以看作是图像-文本对。本文随机抽取60000个图像-文本对进行训练，10000个图像-文本对进行测试，将公开可用的1000维的标签用于体现文本特征。

(4)MIRFLICKR。其包含从Flickr收集的25000个实例，每个实例都是带有相关文本标签的图像。每个图像-文本对都分配有来自38个类别的多个标签。在删除没有文本标签或手动注释标签的图像后，使用数据集中提供的训练集进行检验，用3000维的标签来表示文本特征。

CCA和BLM是两种非监督模型，它们采用成对的信息来最大化投影向量之间的相关性。LCFS，LGCFL，JFSSL是3种有监督的模型，它们利用语义类信息直接将一个模态的数据与另一个模态的数据进行关联。LGCFL方法通过对标签空间的移动，可以扩大类之间的距离，在回归过程中添加对群体的稀疏约束来学习判别式[8]。而JFSSL方法则在投影空间中加入了正则项[3]。

3.2 参数设置

为了更好地与已有的方法进行对比，将WIKI，TVGraz，NUS-WIDE和MIRFLICKR的空间维度分别设置为10，20，10和40。在{0.001，0.01，0.1，1，10}的范围内不断调整参数，分析方法的性能。对于其他的几种方法，根据相应的数据集来设置其参数值大小。

3.3 不同方法的MAP对比结果

平均精度均值(MAP)是用于评估不同方法检索结果的有效性指标。为了验证ECA-CMSA的性能，分别进行两类定向交叉模式检索任务：图像-文本查询和文本-图像查询。如果两类数据点的标签相同，则认为该信息具有一定的相关性。

在WIKI数据集上，将各种方法进行对比，结果如表1所列。由表1可以观察到，本文提出的ECA-CMSA方法的性能有显著提升，其原因可能是ECA-CMSA利用了嵌入矩阵，同时保留了原始特征和语义信息。与CCA和BLM两种无监督方法相比，本文方法提升的效果较好。语义信息提供了模态之间的交互信息和各模态内的信息，而原始特征信息则考虑了各模态之间的相似性。

表1 在WIKI数据集上不同方法的MAP对比结果

Table 1 Compared results of MAP of different methods on WIKI

方法	$R=40$			$R=$全部		
	图像-文本	文本-图像	平均值	图像-文本	文本-图像	平均值
CCA	0.435	0.547	0.491	0.426	0.419	0.423
BLM	0.447	0.537	0.492	0.442	0.439	0.441
LCFS	0.462	0.563	0.513	0.458	0.437	0.448
LGCFL	0.479	0.568	0.524	0.467	0.452	0.460
JFSSL	0.476	0.579	0.528	0.463	0.459	0.461
ECA-CMSA	**0.489**	**0.582**	**0.536**	**0.475**	**0.468**	**0.472**

从表2可以看出，在TVGraz数据集上，ECA-CMSA对于两类检索任务同样达到了最佳效果。对于无监督和有监督方法而言，ECA-CMSA对于图像查询文本任务的性能提升幅度优于文本查询图像。相比其他方法，二者的查询结果均有所提升。

表2 在TVGraz数据集上不同方法的MAP对比结果

Table 2 Compared results of MAP of different methods on TVGraz

方法	$R=40$			$R=$全部		
	图像-文本	文本-图像	平均值	图像-文本	文本-图像	平均值
CCA	0.627	0.614	0.621	0.619	0.608	0.614
BLM	0.634	0.621	0.628	0.621	0.615	0.618
LCFS	0.645	0.637	0.641	0.634	0.626	0.630
LGCFL	0.653	0.646	0.650	0.647	0.637	0.642
JFSSL	0.659	0.642	0.651	0.651	0.648	0.650
ECA-CMSA	**0.664**	**0.658**	**0.661**	**0.661**	**0.652**	**0.657**

表3列出了NUS-WIDE数据集上各种方法的MAP。LGCFL和ECA-CMSA方法比CCA的表现更好，因为都考虑了语义信息。此外，CCA和BLM两种无监督方法的效果优于其他有监督方法，可能是因为NUS-WIDE原始的信息更适用于无监督模式下的还原。NUS-WIDE数据集比WIKI和MIRFLICKR数据集更大，因此语义信息在NUS-WIDE中具有更多的交互性，可以尽可能找到不同模态的数据之间的相似信息。

表3 在NUS-WIDE数据集上不同方法的MAP对比结果

Table 3 Compared results of MAP of different methods on NUS-WIDE

方法	$R=40$			$R=$全部		
	图像-文本	文本-图像	平均值	图像-文本	文本-图像	平均值
CCA	0.772	0.765	0.769	0.765	0.769	0.767
BLM	0.849	0.834	0.842	0.834	0.841	0.838
LCFS	0.841	0.827	0.834	0.839	0.832	0.836
LGCFL	0.785	0.776	0.781	0.773	0.772	0.773
JFSSL	0.769	0.767	0.768	0.757	0.761	0.759
ECA-CMSA	**0.863**	**0.856**	**0.860**	**0.849**	**0.845**	**0.847**

在MIRFLICKR数据集上，由表4可以看出，本文方法的MAP值优于其他方法，而JFSSL的效果次之，CCA和BLM两种无监督方法的效果并不十分理想。ECA-CMSA方法具有既保留原始特征又保留语义信息的能力，学习出了语义标签空间的特征代码向量。由此表明，本文方法和JFSSL方法对于查询有标签的空间信息是有效的。

表 4 在 MIRFLICKR 数据集上不同方法的 MAP 对比结果

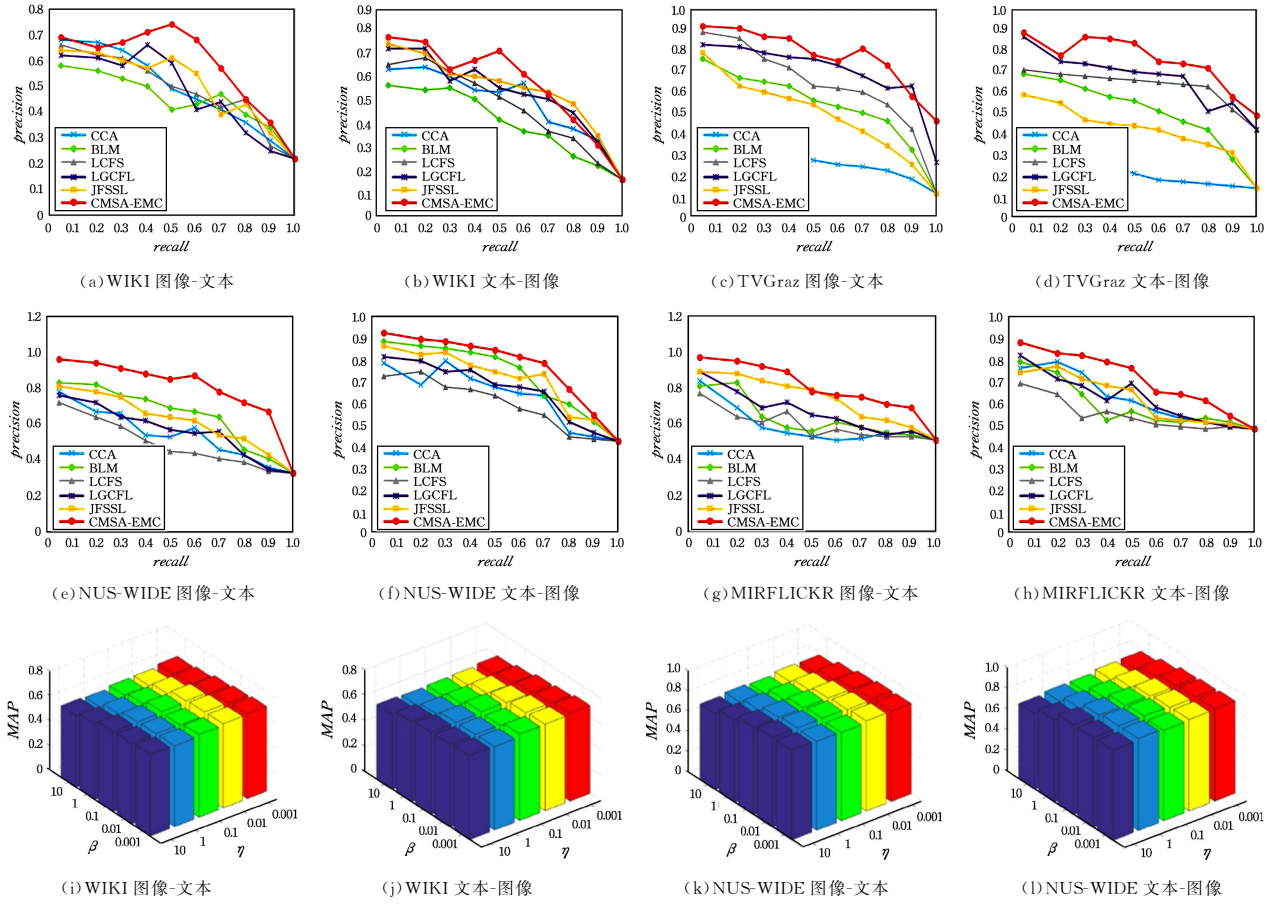
Table 4 Compared results of MAP of different methods on MIRFLICKR

方法	$R=40$			$R=$全部		
	图像-文本	文本-图像	平均值	图像-文本	文本-图像	平均值
CCA	0.871	0.858	0.865	0.734	0.736	0.735
BLM	0.869	0.851	0.860	0.739	0.743	0.741
LCFS	0.896	0.874	0.885	0.754	0.752	0.753
LGCFL	0.894	0.879	0.887	0.762	0.768	0.765
JFSSL	0.903	0.891	0.897	0.773	0.786	0.780
ECA-CMSA	**0.926**	**0.982**	**0.954**	**0.796**	**0.814**	**0.805**

3.4 精确率-召回率曲线和灵敏度分析

由图 2 可知,对于图像-文本查询任务,整体上 ECA-CMSA 的查询效果几乎超过了所有的其他方法。在 MIRFLICKR 数据集上,各种方法的最低精确率均相对较高。在 NUS-WIDE 数据集上,本文方法的性能优势更为明显,相比无监督方法而言,也取得了较好的提升效果。总体而言,ECA-CMSA 提升了图像查询文本任务的性能。

对于文本-图像查询任务,本文方法在 4 个基准数据集上的查全率也比其他方法更高。



(a) WIKI 图像-文本 (b) WIKI 文本-图像 (c) TVGraz 图像-文本 (d) TVGraz 文本-图像

(e) NUS-WIDE 图像-文本 (f) NUS-WIDE 文本-图像 (g) MIRFLICKR 图像-文本 (h) MIRFLICKR 文本-图像

(i) WIKI 图像-文本 (j) WIKI 文本-图像 (k) NUS-WIDE 图像-文本 (l) NUS-WIDE 文本-图像

图 2 精确-召回率曲线和灵敏度分析结果图

Fig. 2 Accuracy-recall rate curve and sensitivity analysis results

在 WIKI 和 NUS-WIDE 数据集上,在{0.001,0.01,0.1,1,10}范围内分别调整两个参数值的大小。当参数变化时,ECA-CMSA 方法的效果也会有所不同,参数范围为 0.001~1 时,该方法可以得到较好的结果。

3.5 收敛性分析

图 3 给出了 ECA-CMSA,CCA 和 BLM 这 3 种无监督方法在两个数据集上进行 10 次以上迭代的收敛损失曲线。可以看出,在 WIKI 和 NUS-WIDE 上,随着迭代次数的增加,3 种方法的损失值不断减少。

经过较少的迭代次数,损失一直变小并趋于稳定,且本文方法的收敛速度最快,因此认为本文方法最终具有一定的收敛性。

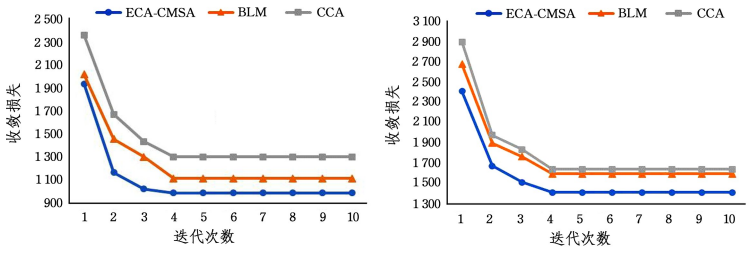


图 3 收敛损失曲线

Fig. 3 Convergence loss curve

结束语 本文提出了一种有效的跨模态检索方法。该方法通过对多模态数据的嵌入映射共识，同时保留了原始特征信息和语义信息，得到了语义代码向量。成对的编码器-解码器是线性对称的，将特征投影返回到原始数据，最小化重构误差。在目标函数中引入平衡参数，并添加正则化稀疏约束。实验结果证明，该自动编码器有效完成了查询任务，提升了检索性能。

跨模态检索技术涉及与数学和统计相关的基础知识，可以应用于与计算机和网络有关的领域，如深度学习和子空间学习。下一步，可以在深度神经网络模型上使用诸如面部表情和身体运动之类的人体特征，在多种模式学习中实现同步。ECA-CMSA 可以还原图像和文本信息之间更多的相似性特征，该模型突破了传统方法的障碍，创新地使用深度学习方法将多模态数据转换为抽象表达，从而可以获得更好的准确性和识别结果。

参 考 文 献

[1] NIE L, ZHAO Y L, AKBARI M, et al. Bridging the vocabulary gap between health seekers and healthcare knowledge[J]. IEEE Trans. Knowl. DataEng., 2015, 27 (2): 396-409.

[2] ABHISHEK S, ABHISHEK K, DAUME H, et al. Generalized multi-view analysis: a discriminative latent space[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2012: 2160-2167.

[3] WANG K, HE R, WANG L, et al. Joint feature selection and subspace learning for cross-modal retrieval[J]. Trans Pattern Anal. Mach. Intell., 2016, 38: 2010-2023.

[4] PUTTHIVIDHY D, ATTIAS H T, NAGARAJAN S S. Topic regression multi-modal latent dirichlet allocation for image annotation[C]// Proceedings of Conference on Computer Vision and Pattern Recognition. CVPR, 2010.

[5] MUNOZ L, RAMOS J. Similarity-based Heterogeneous Neural Networks[J]. Engineering Letters, 2007, 14(2): 103-116.

[6] ZHOU J, DING G, GUO Y. Latent semantic sparse hashing for cross-modal similarity search[C]// Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. ACM, 2014: 415-424.

[7] WU Y L, WANG S H, HUANG Q M. Multi-modal semantic autoencoder for cross-modal retrieval[J]. Neurocomputing, 2019, 331: 165-175.

[8] KANG C, XIANG S, LIAO S, et al. Learning consistent feature representation for cross-modal multimedia retrieval[J]. IEEE Trans. Multimed., 2015, 17(3): 370-381.

[9] DAI X M, LI S G. Cross-modal deep discriminant analysis[J]. Neurocomputing, 2018, 314: 437-444.

[10] 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.

[11] YANG W, YI D, LEI Z, et al. 2d-3d face matching using cca [C]// Proceedings of the 8th IEEE International Conference on Automatic Face & Gesture Recognition (FG'08). IEEE, 2008: 1-6.

[12] SHARMA A, JACOBS D W. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch[C]// Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2011: 593-600.

[13] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders [C]// Proceedings of the 25th International Conference on Machine Learning. ACM, 2008: 1096-1103.

[14] LANGE S, RIEDMILLER M. Deep Auto-Encoder Neural Networks in Reinforcement Learning[C]// International Joint Conference on Neural Networks (IJCNN 2010). Barcelona, Spain, 2010: 18-23.

[15] SAINATH T N, KINGSBUR Y B, RAMABHADRAN B. Auto-encoderbottleneck features using deep belief networks[C]// 2012 IEEE International Conference on IEEE Acoustics, Speech and Signal Processing (ICASSP). 2012: 4153-4156.

[16] ZHANG L, MA B, LI G, et al. PL-ranking: a novel ranking method for cross-modal retrieval[C]// Proceedings of the ACM on Multimedia Conference. ACM, 2016: 1355-1364.

[17] PEREIRA J C, COVIELLO E, DOYLE G, et al. On the role of correlation and abstraction in cross-modal multimedia retrieval [J]. TPAMI, 2014, 36(3): 521-535.

[18] 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 CIVR. ACM, 2009: 48.

[19] HUISKES M J, LEW M S. The mirflickr retrieval evaluation [C]// Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval. ACM, 2008: 39-43.

SUN Sheng-zi, born in 1996, postgraduate, is a member of China Computer Federation. Her main research interests include artificial intelligence and pattern recognition.

GUO Bing-hui, born in 1982, associate professor, is a professional member of China Computer Federation. His main research interests include data science and complex intelligent system.