计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202220430
60(11): 2660−2670, 2023

面向科技资讯的基于语义对抗和媒体对抗的跨媒体检索方法

李 昂 杜军平 寇菲菲 薛 哲 徐 欣 许明英 姜 阳
(北京邮电大学计算机学院(国家示范性软件学院) 北京 100876)
(智能通信软件与多媒体北京市重点实验室(北京邮电大学) 北京 100876)
(junpingdu@126.com)

Scientific and Technological Information Oriented Semantics-Adversarial and Media-Adversarial Based Cross-Media Retrieval Method

Li Ang, Du Junping, Kou Feifei, Xue Zhe, Xu Xin, Xu Mingying, and Jiang Yang
(*School of Computer Science* (*National Pilot School of Software Engineering*), *Beijing University of Posts and Telecommunications*, *Beijing* 100876)
(*Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia* (*Beijing University of Posts and Telecommunications*) *Beijing* 100876)

Abstract Cross-media retrieval of scientific and technological information is one of the important tasks in the cross-media study. Cross-media scientific and technological information retrieval obtains target information from massive multi-source and heterogeneous scientific and technological resources, which helps to design applications that meet users' needs, including scientific and technological information recommendation, personalized scientific and technological information retrieval, etc. The core of cross-media retrieval is to learn a common subspace, in which data from different media can be directly compared with each other. In subspace learning, existing methods often focus on modeling the discrimination of intra-media data and the invariance of inter-media data after mapping, while ignoring semantic consistency within media and media discrimination within semantics, which limits the result of cross-media retrieval. In light of this, we propose a scientific and technological information oriented semantics-adversarial and media-adversarial cross-media retrieval method (SMCR) to find an effective common subspace. Specifically, SMCR minimizes the loss of inter-media semantic consistency in addition to modeling intra-media semantic discrimination, to preserve semantic similarity before and after mapping. Furthermore, SMCR constructs a basic feature mapping network and a refined feature mapping network to jointly minimize the media discriminative loss within semantics, to enhance the feature mapping network's ability to confuse the media discriminant network. Experimental results on two datasets demonstrate that the proposed SMCR outperforms state-of-the-art methods in cross-media retrieval.

Key words cross-media retrieval; adversarial learning; scientific and technological information; media constraint; semantic consistency

摘 要 科技资讯跨媒体检索是跨媒体领域的重要任务之一,面临着多媒体数据间异构鸿沟和语义鸿沟亟待打破的难题.通过跨媒体科技资讯检索,用户能够从多源异构的海量科技资源中获取目标科技资讯.

收稿日期: 2022-05-28; **修回日期**: 2022-11-18
基金项目: 国家自然科学基金重大项目(62192784); 第八届青年人才托举工程项目(2022QNRC001)
This work was supported by the Major Program of the National Natural Science Foundation of China (62192784) and the 8th Young Elite Scientists Sponsorship Program by CAST (2022QNRC001).
通信作者: 杜军平(junpingdu@126.com)

这有助于设计出符合用户需求的应用，包括科技资讯推荐、个性化科技资讯检索等. 跨媒体检索研究的核心是学习一个公共子空间，使得不同媒体的数据在该子空间中可以直接相互比较. 在子空间学习中，现有方法往往聚焦于建模媒体内数据的判别性和媒体间数据在映射后的不变性，却忽略了媒体间数据在映射前后的语义一致性和语义内的媒体判别性，使得跨媒体检索效果存在局限性. 鉴于此，提出一种面向科技资讯的基于语义对抗和媒体对抗的跨媒体检索方法（SMCR），寻找可供映射的有效公共子空间. 具体而言，SMCR 在建模媒体内语义判别性之外，将媒体间语义一致性损失最小化，以保留映射前后的语义相似性. 此外，SMCR 构建基础特征映射网络和精炼特征映射网络，联合最小化语义内的媒体判别性损失，有效增强了特征映射网络混淆媒体判别网络的能力. 在 2 个数据集上的大量实验结果表明，所提出的 SMCR 方法在跨媒体检索中的表现优于最前沿的方法.

关键词 跨媒体检索；对抗学习；科技资讯；媒体约束；语义一致性

中图法分类号 TP391

科技资讯聚焦了中外高新技术的前沿动态. 实时跟进最新的科技资讯，有助于促进国家战略科技力量的发展，驱动科技创新，进而确保国家高质量发展[1]. 科技资讯中包含大量的多媒体信息（如图像、文本等），具备体量大、来源丰富、类型多样等特点[2-3]. 随着用户感兴趣的科技资讯模态不再单一，检索需求也呈现出从单一模态到跨模态的发展态势[4-5]. 凭借跨媒体科技资讯检索，用户能够从多源异构的海量科技资源中获取目标科技资讯；研究者亦能近一步设计出符合用户需求的应用，包括科技资讯推荐[6]、个性化科技资讯检索[7] 等. 跨媒体科技资讯检索作为当下的研究热点，仍旧面临着多媒体数据间异构鸿沟和语义鸿沟亟待打破的难题[8-9]. 本文旨在解决现有跨媒体科技资讯检索中仅考虑了媒体内数据判别损失和媒体间数据在映射后的不变性损失，却忽略了媒体间数据在映射前后的语义一致性损失和语义内的媒体判别性损失，使得跨媒体检索效果存在局限性的问题.

跨媒体科技资讯检索方法种类繁多. 先前的工作[10-14] 聚焦于传统的统计关联分析方法，通过优化统计值来学习公共空间的线性投影矩阵[15]，目的是建立一个共享子空间，使得不同媒体类型的数据对象的相似性可以映射到该子空间中，再使用常见的距离进行度量. 然而，文献 [10−14] 所述的方法依赖于数据的线性表示，仅通过线性投影很难完全模拟现实世界中跨媒体数据的复杂相关性. 因此，一些研究[16-20] 通过深度学习方法解决上述问题，利用其强大的抽象能力处理多媒体数据的多层非线性变换，进行跨媒体相关学习. 然而，现有的基于深度学习的跨媒体检索模型通常只专注于保留耦合的跨媒体样本（例如图像和文本）的成对相似性[21]，却忽略了一种媒体的一个样本可能存在多个相同媒体的语义不同的样本，因此无法保留跨媒体语义结构. 保留跨媒体语义结构需要使得相同语义不同媒体的数据间距离最小化，且相同媒体不同语义的数据间距离最大化. 最近的工作[22-26] 引入对抗学习的思想，通过联合执行标签预测并保留数据中的底层跨媒体语义结构，为公共子空间中不同媒体的样本生成媒体不变表示. 然而，文献 [22−26] 所述的方法聚焦于建模媒体内数据的语义判别性和媒体间数据在子空间映射后的语义不变性，却忽略了媒体间数据在映射前后的语义一致性和语义内的媒体判别性，使得跨媒体检索效果存在局限性.

针对上述问题，引入语义内的媒体约束来加强将不同类型的媒体数据映射到共享高级语义空间的能力，提出一种面向科技资讯的基于语义对抗和媒体对抗的跨媒体检索（semantics-adversarial and media-adversarial cross-media retrieval, SMCR）方法. SMCR 方法采用对抗博弈[27] 的思想，构建特征映射器和媒体判别器，进行极小化极大化游戏.SMCR 方法追随先前工作[28-29]，采用标签预测来确保数据在特征投影后仍保留在媒体内的区别. 与先前工作不同的是，SMCR 方法同时最小化相同语义的文本−图像对中不同媒体的数据分别在特征映射前和特征映射后的距离，以确保不同媒体间数据在映射过程中的语义一致性得以保留. 此外，通过构建基础映射网络和精炼映射网络共同辅助建模语义内的媒体约束，使映射后的数据做到语义上接近自身和媒体上远离自身，来增强特征映射网络混淆媒体判别网络的能力. 媒体判别网络负责区分数据的原始媒体，一旦媒体判别网络被欺骗，整个博弈过程收敛.

本文的主要贡献包括 3 个方面：

1）提出一种面向科技资讯的基于语义对抗和媒体对抗的跨媒体检索方法（SMCR），通过端到端的方

式同时保持媒体内的语义判别性、媒体间的语义一致性、语义内的媒体判别性，能够有效地学习异构数据的公共表示；

2)通过构建基础特征映射网络和精炼特征映射网络联合进行多媒体数据特征映射，辅助语义内的媒体约束，有效地增强了特征映射网络混淆媒体判别网络的能力；

3) 在 2 个数据集上进行的大量实验表明，本文提出的 SMCR 方法优于当前最前沿的跨媒体检索方法，包括传统的方法和基于深度学习的方法.

1 相关工作

科技资讯跨媒体检索是近年来的研究热点，旨在学习一个公共子空间[13,24,30]，使得不同媒体的数据在该子空间中可以直接相互比较，以跨越不同媒体间存在的语义鸿沟.

一类经典的方法当属传统的统计关联分析方法[10-14]，它是公共空间学习方法的基本范式和基础，主要通过优化统计值来学习公共空间的线性投影矩阵. 例如，Hardoon 等人[12]提出典型关联分析(canonical correlation analysis, CCA)方法，CCA 方法是一种关联 2 个多维变量之间线性关系的方法，可以被视为使用复杂标签作为引导特征选择朝向底层语义的一种方式. 该方法利用同一语义对象的 2 个视角来提取语义的表示. Wang 等人[13]提出一种基于耦合特征选择和子空间学习的联合学习(joint feature selection and subspace learning, JFSSL)，受 CCA 和线性最小二乘法之间潜在关系的启发，将耦合线性回归用于学习投影矩阵，使来自不同媒体的数据映射到公共子空间中. 同时，JFSSL 将l_2正则用于同时从不同的特征空间中选择相关和不相关的特征，并且在映射时使用多媒体图正则化来保留媒体间和媒体内的相似性关系.Zhai 等人[14]提出了一种新的跨媒体数据特征学习算法，称为联合表示学习(joint representation learning, JRL). 该方法能够在统一的优化框架中联合探索相关性和语义信息，并将所有媒体类型的稀疏和半监督正则化集成到一个统一的优化问题中. JRL 旨在同时学习不同媒体的稀疏投影矩阵，并将原始异构特征直接投影到联合空间中. 然而，仅通过线性投影很难完全模拟现实世界中跨媒体数据的复杂相关性.

随着深度学习的兴起，许多研究聚焦于将能够实现多层非线性变换的深度神经网络应用于跨媒体检索中[16-20]. 例如，Yan 等人[17]提出一种基于深度典型相关分析(deep canonical correlation analysis, DCCA)的跨媒体图像字幕匹配方法. 通过解决非平凡的复杂性和过度拟合问题，使该方法适用于高维图像和文本表示以及大型数据集.Peng 等人[18]提出一种跨媒体多重深度网络(cross-media multiple deep network, CMDN)，通过分层学习来利用复杂而丰富的跨媒体相关性. 在第 1 阶段，CMDN 不像先前工作仅利用媒体内的分离表示，而是联合学习每种媒体类型的 2 种互补的分离表示；在第 2 阶段，由于每种媒体类型都有 2 个互补的独立表示，该方法在更深的 2 级网络中分层组合单独的表示，以便联合建模媒体间和媒体内的信息以生成共享表示. 然而，现有的基于深度神经网络的跨媒体检索模型通常只专注于保留耦合的跨媒体样本(例如图像和文本)的成对相似性，却忽略了一种媒体的一个样本，可能存在多个相同媒体的语义不同的样本，因此无法保留跨媒体语义结构.

近年来，相关研究转而向对抗学习[31]进行探索. 虽然它在图像生成[32]中应用较广，但研究者也将其用作正则化器[33]. 一些研究将其思想应用于跨媒体检索，并取得了显著的效果[22-26]. 例如，Wang 等人[24]提出一种基于对抗跨媒体检索(adversarial cross-modal retrieval, ACMR)方法来解决跨媒体语义结构难保留的问题. 该方法使用特征投影器，通过联合执行标签预测并保留数据中的底层跨媒体语义结构，为公共子空间中不同媒体的样本生成媒体不变表示. ACMR 的目的是混淆充当对手的媒体分类器，媒体分类器试图根据它们的媒体来区分样本，并以这种方式引导特征投影器的学习. 通过这个过程的收敛，即当媒体分类器失败时，表示子空间对于跨媒体检索是最优的. Zhen 等人[25]提出一种深度监督跨媒体检索(deep supervised cross-modal retrieval, DSCMR)方法，旨在找到一个共同的表示空间，以便在其中直接比较来自不同媒体的样本. 该方法将标签空间和公共表示空间中的判别损失最小化，以监督模型学习判别特征. 同时最小化媒体不变性损失，并使用权重共享策略来消除公共表示空间中多媒体数据的跨媒体差异，以学习媒体不变特征. 刘翀等人[26]提出一种基于对抗学习和语义相似度的社交网络跨媒体搜索方法(semantic similarity based adversarial cross media retrieval, SSACR)，SSACR 使用语义分布及相似度作为特征映射网训练依据，使得相同语义下的不同媒体数据在该空间距离小、不同语义下的相同媒体数

据距离大，最终在同一空间内使用相似度来排序并得到搜索结果. 然而，文献 [24−26] 聚焦于建模媒体内数据语义损失和媒体间数据在映射后的语义损失，却忽略了媒体间数据在映射前后的语义一致性和语义内的媒体判别性，使得跨媒体检索效果存在局限性.

2 问题定义

多媒体数据种类繁多，为了不失通用性，本文聚焦于文本、图像 2 种媒体的跨媒体检索. 给定一系列语义相关的图像-文本对 $m=\{m_1, m_2, \cdots, m_{|m|}\}$，其中 $m_i=(\boldsymbol{v}_i, \boldsymbol{t}_i)$表示 m中的第 i 个图像-文本对，$\boldsymbol{v}_i \in \mathbb{R}^{d_{\text{vis}}}$表示维度为$d_{\text{vis}}$的图像特征向量，$\boldsymbol{t}_i \in \mathbb{R}^{d_{\text{tex}}}$表示维度为$d_{\text{tex}}$的文本特征向量. 每个图像-文本对都对应着一个语义类别向量$\boldsymbol{l}_i=(y_1, y_2, \cdots, y_C) \in \mathbb{R}^C$，用来表示图像-文本对的语义分布，也可以表示类别标签分布. 其中 C 表示语义类别总数，假设$\boldsymbol{l}_i$属于第j个语义类别，则记 $y_j=1$，否则记$y_j=0$. 记m 中所有的图像、文本、语义类别所对应的特征矩阵为$\boldsymbol{V}=(\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N) \in \mathbb{R}^{d_{\text{vis}} \times N}$，$\boldsymbol{T}=(\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_N) \in \mathbb{R}^{d_{\text{tex}} \times N}$，$\boldsymbol{L}=(\boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_N) \in \mathbb{R}^{C \times N}$.

我们的目标是利用一种媒体的数据（如图像$\boldsymbol{v}_i$或文本$\boldsymbol{t}_i$）检索另一种媒体的数据（如文本$\boldsymbol{t}_i$或图像$\boldsymbol{v}_i$）. 为了比较不同媒体数据之间的语义相似性，我们设计 2 个特征映射网络——基础映射网络和精炼映射网络. 基础映射网络将图像特征和文本特征映射到统一的隐语义空间S中以进行语义相似性的对比. 图像特征$\boldsymbol{V}$映射到隐语义空间S后的特征记为$S_V=f_V(\boldsymbol{V};\theta_V)$，文本特征$\boldsymbol{T}$映射到隐语义空间$S$后的特征记为$S_T=f_T(\boldsymbol{T};\theta_T)$. 其中 $f_V(\boldsymbol{V};\theta_V)$和 $f_T(\boldsymbol{T};\theta_T)$分别表示图像和文本的映射函数. 为了近一步提高特征映射质量，我们用精炼映射网络对基础映射网络的输出特征进行映射. 图像特征S_V映射后的特征记为$S'_V=g_{S_V}(S_V;\theta_{S_V})$，文本特征$S_T$映射后的特征记为$S'_T=g_{S_T}(S_T;\theta_{S_T})$. 其中$g_{S_V}(S_V;\theta_{S_V})$和$g_{S_T}(S_T;\theta_{S_T})$表示图像特征和文本特征的映射函数.

3 面向科技资讯的基于语义对抗和媒体对抗的跨媒体检索方法

本文提出一种面向科技资讯的基于语义对抗和媒体对抗的跨媒体检索方法（SMCR）.SMCR 的框架如图 1 所示. 本文的目的是利用对抗学习的思想不断在语义与媒体间进行对抗，学习到一个公共子空间，使不同媒体的数据在该子空间中可以直接相互比较.

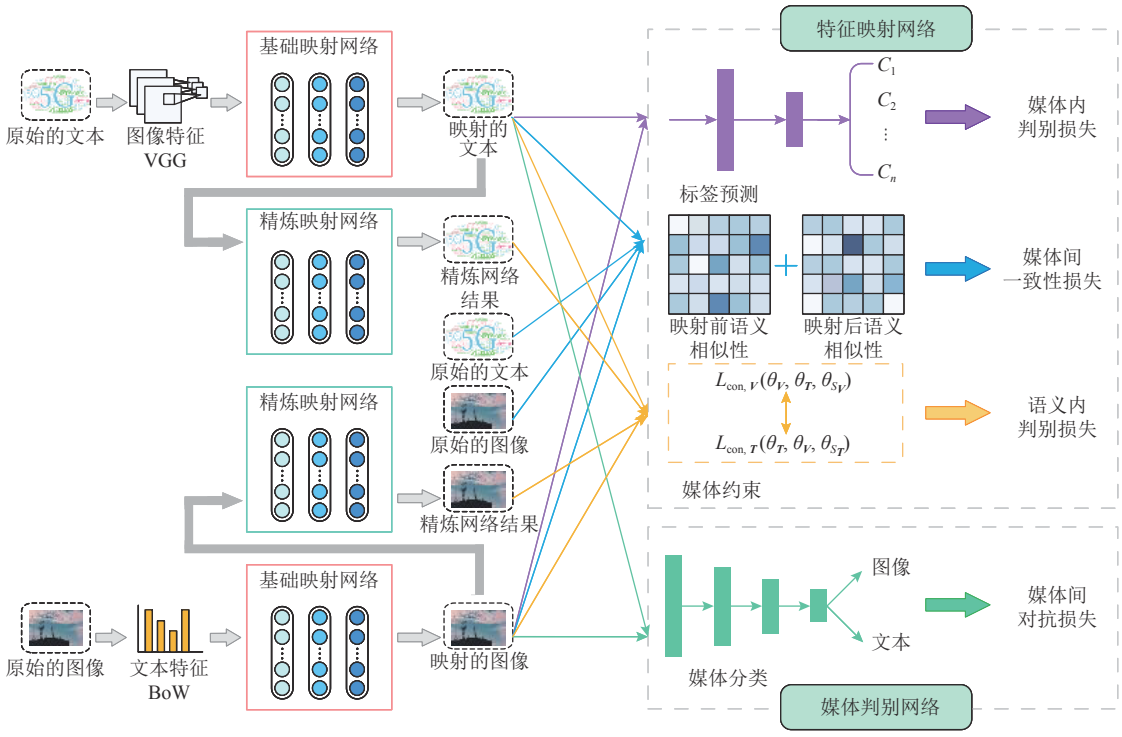


Fig. 1 The overall framework of SMCR

图 1 SMCR 的整体框架

3.1 特征映射网络

本文采用特征映射网络是为了将不同媒体的特征映射到统一的隐语义空间以便进行语义相似性的比较. 同时, 特征映射网络也扮演着 GAN[27] 中“生成器”的角色, 目的是为了迷惑媒体判别网络(将在 3.2 节介绍). 为了使映射后的特征表示充分考虑 2 类媒体数据的语义相似性和媒体相似性, 本文设计的特征映射网络由 3 部分组成: 媒体内的标签预测、媒体间的语义保留、语义内的媒体约束. 媒体内的标签预测使得映射在隐语义空间S中的特征依然能够以原始的语义标签为真值进行语义分类; 媒体间的语义保留使得语义相同媒体不同的数据在映射前后都能保留语义相似性; 语义内的媒体约束使得映射后的数据更加逼近原本语义.

3.1.1 标签预测

为了保证映射到隐语义空间S中的特征依然能够保留原始语义, 以原始的语义标签为真值进行语义分类. 在每个特征映射网络的最后加入一个保持线性激活的 softmax 层. 将图像-文本对$m_i=(\boldsymbol{v}_i,\boldsymbol{t}_i)$作为样本进行训练, 并输出每个数据对应语义类别的概率分布. 采用在文献 [24] 中介绍的损失函数来计算媒体内的判别损失:

$$L_{\text{imd}}(\theta_{\text{imd}})=-\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{l}_i\cdot(\ln\hat{p}_i(\boldsymbol{v}_i)+\ln\hat{p}_i(\boldsymbol{t}_i))). \tag{1}$$

其中L_{imd}表示对所有图像-文本对进行语义类别分类的交叉熵损失, θ_{imd}表示分类器的参数, $\boldsymbol{l}_i$是每个样本m_i的真值, $\hat{p}_i$是样本中每个数据(图像或文本)所得到的概率分布.

3.1.2 语义保留

语义保留模块致力于保证语义相同、媒体不同的数据在映射前后都能保留语义相似性, 即媒体不同、语义相同的数据距离较近, 媒体不同、语义不同的数据距离较远. 在映射到隐语义空间S之前, 每个样本m_i中的图像数据与文本数据的语义分布分别为$\boldsymbol{l}_{\text{vis}}$和$\boldsymbol{l}_{\text{tex}}$, 那么 2 个不同媒体数据间的语义一致性损失用$l_2$范数表示为

$$l_2(\boldsymbol{l}_{\text{vis}},\boldsymbol{l}_{\text{tex}})=\|\boldsymbol{l}_{\text{vis}}-\boldsymbol{l}_{\text{tex}}\|_2. \tag{2}$$

在映射到隐语义空间S之后, 每个样本m_i中的图像数据特征S_V与文本数据的特征S_T之间的语义一致性损失同样用l_2范数表示为

$$l_2(S_V,S_T)=\|f_V(\boldsymbol{V};\theta_V)-f_T(\boldsymbol{T};\theta_T)\|_2. \tag{3}$$

因此, 整体的媒体间一致性损失可以建模为$l_2(\boldsymbol{l}_{\text{vis}},\boldsymbol{l}_{\text{tex}})$和$l_2(S_V,S_T)$两者的结合:

$$L_{\text{imi}}(\theta_V,\theta_T)=l_2(\boldsymbol{l}_{\text{vis}},\boldsymbol{l}_{\text{tex}})+l_2(S_V,S_T), \tag{4}$$

其中L_{imi}表示媒体间同时考虑映射前与映射后的语义一致性损失.

3.1.3 媒体约束

除了便于度量不同媒体数据间的语义相似性之外, 特征映射网络的另一个作用是生成映射后的特征来欺骗媒体判别网络, 让它无法区分出数据的原始媒体. 因此, 引入语义内的媒体约束模块. 为了能够更加逼真地映射出难以区分媒体的特征, 在基础的特征映射网络P_1之外, 构造另一个相同结构的特征映射网络P_2, 称为精炼网络. 精炼网络P_2的输入是P_1的输出结果S_V或S_T. P_2的输出是$S'_V=g_{S_V}(S_V;\theta_{S_V})$或$S'_T=g_{S_T}(S_T;\theta_{S_T})$. 其中$S'_V$和$S'_T$分别表示$S_V$和$S_T$经过特征映射网络$P_2$映射后的特征, $g_{S_V}(S_V;\theta_{S_V})$和$g_{S_T}(S_T;\theta_{S_T})$分别表示$S_V$和$S_T$这 2 种特征的映射函数.

对每一个图像-文本对m_i而言, 目标是让精炼网络P_2映射出的特征(S'_V或S'_T)距离基础网络P_1映射的特征(S_V或S_T)较远, 距离相同语义的特征(S_T或S_V)较近. 受到文献 [34−36] 启发, 语义内的媒体判别损失采用如下约束损失进行计算:

$$L_{\text{con},V}(\theta_V,\theta_T,\theta_{S_V})=\max\left(0,\left\|S'_V-S_T\right\|_2-\left\|S'_V-S_V\right\|_2\right), \tag{5}$$

$$L_{\text{con},T}(\theta_T,\theta_V,\theta_{S_T})=\max\left(0,\left\|S'_T-S_V\right\|_2-\left\|S'_T-S_T\right\|_2\right). \tag{6}$$

其中$L_{\text{con},V}$表示图像媒体数据的约束损失, $L_{\text{con},T}$表示文本媒体数据的约束损失.

因此, 整体语义内的媒体判别损失可以建模为图像媒体数据的约束损失$L_{\text{con},V}(\theta_V,\theta_T,\theta_{S_V})$与文本媒体数据的约束损失$L_{\text{con},T}(\theta_T,\theta_V,\theta_{S_T})$的结合:

$$L_{\text{con}}(\theta_V,\theta_T,\theta_{S_V},\theta_{S_T})=L_{\text{con},V}(\theta_V,\theta_T,\theta_{S_V})+L_{\text{con},T}(\theta_T,\theta_V,\theta_{S_T}). \tag{7}$$

3.1.4 特征映射网络损失

整个特征映射网络的映射性损失由媒体内的判别损失L_{imd}、媒体间的一致性损失L_{imi}、语义内的判别损失L_{com}共同组成, 记为L_{emb}:

$$L_{\text{emb}}(\theta_V,\theta_T,\theta_{S_V},\theta_{S_T},\theta_{\text{imd}})=\alpha\cdot L_{\text{imi}}+\beta\cdot L_{\text{con}}+L_{\text{imd}}, \tag{8}$$

其中α和β为可调节参数, 用以控制L_{imi}和L_{con}这 2 类损失在整个特征映射网络损失中的参与度.

3.2 媒体判别网络

媒体判别网络扮演着 GAN[27] 中“判别器”的角色, 用来判断映射到隐语义空间后的数据的原始媒体. 令经过图像映射函数的数据标签为 0, 经过文本映射函数的数据标签为 1.本文使用一个参数为θ_{dis}的 3 层全连接网络作为判别网络, 充当特征映射网络的对手. 其目标是最小化媒体分类损失, 也称为对抗性

损失L_{adv}，定义为

$$L_{\text{adv}}(\theta_{\text{dis}}) = -\frac{1}{n}\sum_{i=1}^{n}(\ln D(\boldsymbol{v}_i;\theta_{\text{dis}}) + \ln(1-D(\boldsymbol{t}_i;\theta_{\text{dis}}))), \quad (9)$$

其中L_{adv}表示媒体判别网络中每个样本m_i的交叉熵损失，$D(\cdot;\theta_{\text{dis}})$表示样本中每个数据（图像或文本）所得到的媒体概率分布.

3.3 对抗学习

对抗学习的目的旨在通过同时最小化式(8)的映射性损失和式(9)的对抗性损失，来学习得到最优的特征表示网络参数，定义如下所示：

$$(\theta_{\boldsymbol{V}}, \theta_{\boldsymbol{T}}, \theta_{S_{\boldsymbol{V}}}, \theta_{S_{\boldsymbol{T}}}, \theta_{\text{imd}}) = \arg\min_{\theta_{\boldsymbol{V}},\theta_{\boldsymbol{T}},\theta_{S_{\boldsymbol{V}}},\theta_{S_{\boldsymbol{T}}},\theta_{\text{imd}}} (L_{\text{emb}}(\theta_{\boldsymbol{V}}, \theta_{\boldsymbol{T}}, \theta_{S_{\boldsymbol{V}}}, \theta_{S_{\boldsymbol{T}}}, \theta_{\text{imd}}) - L_{\text{adv}}(\theta_{\text{dis}})), \quad (10)$$

$$\theta_{\text{dis}} = \arg\max_{\theta_{\text{dis}}}(L_{\text{emb}}(\theta_{\boldsymbol{V}}, \theta_{\boldsymbol{T}}, \theta_{S_{\boldsymbol{V}}}, \theta_{S_{\boldsymbol{T}}}, \theta_{\text{imd}}) - L_{\text{adv}}(\theta_{\text{dis}})). \quad (11)$$

具体的对抗学习训练过程如算法 1 所示.

算法 1. SMCR 的对抗训练过程.

输入：图像特征矩阵$\boldsymbol{V}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_N)$，文本特征矩阵$\boldsymbol{T}=(\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_N)$，真值语义标签矩阵$\boldsymbol{L}=(\boldsymbol{l}_1,\boldsymbol{l}_2,\cdots,\boldsymbol{l}_N)$，迭代次数$k$，学习率$\mu$，每个批次的数据量$m$，损失参数$\lambda$；

输出：参数$\theta_{\boldsymbol{V}},\theta_{\boldsymbol{T}},\theta_{S_{\boldsymbol{V}}},\theta_{S_{\boldsymbol{T}}}$.

①随机初始化模型参数；

②while 未收敛 do

③ for *iter* =1 to k do

④ 通过随机梯度下降更新参数$\theta_{\boldsymbol{V}},\theta_{\boldsymbol{T}},\theta_{S_{\boldsymbol{V}}},\theta_{S_{\boldsymbol{T}}},\theta_{\text{imd}}$；

⑤ $\theta_{\boldsymbol{V}} \leftarrow \theta_{\boldsymbol{V}} - \mu\cdot\nabla_{\theta_{\boldsymbol{V}}}\frac{1}{m}(L_{\text{emb}}-L_{\text{adv}})$；

⑥ $\theta_{\boldsymbol{T}} \leftarrow \theta_{\boldsymbol{T}} - \mu\cdot\nabla_{\theta_{\boldsymbol{T}}}\frac{1}{m}(L_{\text{emb}}-L_{\text{adv}})$；

⑦ $\theta_{S_{\boldsymbol{V}}} \leftarrow \theta_{S_{\boldsymbol{V}}} - \mu\cdot\nabla_{\theta_{S_{\boldsymbol{V}}}}\frac{1}{m}(L_{\text{emb}}-L_{\text{adv}})$；

⑧ $\theta_{S_{\boldsymbol{T}}} \leftarrow \theta_{S_{\boldsymbol{T}}} - \mu\cdot\nabla_{\theta_{S_{\boldsymbol{T}}}}\frac{1}{m}(L_{\text{emb}}-L_{\text{adv}})$；

⑨ $\theta_{\text{imd}} \leftarrow \theta_{\text{imd}} - \mu\cdot\nabla_{\theta_{\text{imd}}}\frac{1}{m}(L_{\text{emb}}-L_{\text{adv}})$；

⑩ end for

⑪end while

⑫通过随机梯度上升更新参数θ_{dis}；

⑬$\theta_{\text{dis}} \leftarrow \theta_{\text{dis}} + \mu\cdot\lambda\cdot\nabla_{\theta_{\text{dis}}}\frac{1}{m}(L_{\text{emb}}-L_{\text{adv}})$.

4 实验设置

本文分别阐述对实验部分至关重要的研究问题、数据集、对比算法、评价指标等 4 个方面.

4.1 研究问题

本文通过 3 个研究问题来引导实验的设置.

研究问题 1. 面向科技资讯的基于语义对抗和媒体对抗的跨媒体检索方法 SMCR 的表现能否优于前沿的跨媒体检索算法.

研究问题 2. SMCR 方法的主要组成部分对于跨媒体检索是否存在贡献.

研究问题 3. SMCR 方法是否对参数敏感.

4.2 数据集

为了回答上述 3 个研究问题，使用爬取自科技资讯网站 SciTechDaily[37] 的数据集进行实验. 数据集包括 5 217 个图像-文本对，将其中的 4 173 对数据作为训练集，1 044 对数据作为测试集. 为了验证本文模型的通用性，同时使用 Wikipedia[38] 数据集进行实验. Wikipedia 数据集包括 2 866 个图像-文本对，将其中的 2 292 对数据作为训练集，574 对数据作为测试集. 这 2 个数据集的详细信息如表 1 所示.

Table 1 Attributes of Two Datasets Used for the Experiments

表 1 实验使用的 2 个数据集的属性

数据集	训练样本数/测试样本数	标签数	图像特征	文本特征
SciTechDaily	4 173/1 044	8	4096d VGG	6500d BoW
Wikipedia	2 292/574	10	4096d VGG	5000d BoW

4.3 对比算法

本文将 SMCR 与相关的基准算法和前沿算法进行比较，对比算法如下.

1）典型关联分析（canonical correlation analysis, CCA）. 该模型[12] 为不同的媒体类型的数据学习一个公共子空间，使 2 组异构数据之间的关联最大化.

2）基于耦合特征选择和子空间学习的联合学习（joint feature selection and subspace learning, JFSSL）. 该模型[13] 学习投影矩阵将多媒体数据映射到一个公共子空间，并同时从不同的特征空间中选择相关的和有区别的特征.

3）跨媒体多重深度网络（cross-media multiple deep network, CMDN）. 该模型[18] 通过分层学习来利用复杂的跨媒体相关性. 在第 1 阶段，联合对媒体内和媒体信息进行建模；在第 2 阶段，分层组合媒体间表示和媒体内表示来进一步学习丰富的跨媒体相关性.

4）基于对抗的跨媒体检索（adversarial cross-modal retrieval, ACMR）. 该模型[24] 基于对抗性学习寻求有效的公共子空间. 对特征投影器施加 3 重约束，以最小化来自具有相同语义标签、不同媒体的所有

样本表示之间的差距, 同时最大化语义不同的图像和文本之间的距离.

5)深度监督跨媒体检索(deep supervised cross-modal retrieval, DSCMR). 该模型[25]同样基于对抗性学习的思想, 将标签空间和公共表示空间中的判别损失最小化, 同时最小化媒体不变性损失, 并使用权重共享策略来消除公共表示空间中多媒体数据的跨媒体差异.

6)基于对抗学习和语义相似度的社交网络跨媒体搜索(SSACR). 该模型[26]同样基于对抗性学习的思想, 将映射到同一语义空间的不同媒体数据的特征向量进行了相似度计算, 并与原本的语义特征向量之间的相似度进行比较, 以消除同一语义下不同媒体数据的差异.

4.4 评价指标

本文采用跨媒体检索[39-40]中经典的评价指标——平均精度均值(mean average precision, *mAP*), 在文本检索图像 txt2img 和图像检索文本 img2txt 这 2 个任务上, 分别对 SMCR 和所有对比算法进行评价. 计算 *mAP*, 首先需计算 R个检索出的文档的平均精度 $AP=\frac{1}{T}\sum_{r=1}^{R}P(r)\delta(r)$. 其中$T$是检索出的文档中的相关文档数量, $P(r)$表示前r个检索出的文档的精度, 如果第r个检索出的文档是相关的, 则 $\delta(r)=1$, 否则 $\delta(r)=0$. 然后通过对查询集中所有查询的 *AP* 值进行平均来计算 *mAP*. *mAP* 值越大, 说明跨媒体检索结果越精准.

5 实验结果与分析

本节对所有实验结果进行分析, 来回答 4.1 节提出的研究问题.

5.1 SMCR 算法的有效性

为了回答研究问题 1, 将 SMCR 和 6 个前沿算法分别在 SciTechDaily, Wikipedia 这 2 个数据集上进行对比. 对比算法为: 1)基于统计关联分析的方法 CCA[12], JFSSL[13]; 2)基于深度学习的方法 CMDN[18], ACMR[24], DSCMR[25], SSACR[26].

表 2 展示了本文在文本检索图像 txt2img 和图像检索文本 img2txt 这 2 个任务上, 对前 5 个、前 25 个、前 50 个的检索结果计算 *mAP* 值(*mAP*@5, *mAP*@25, *mAP*@50)和 2 个检索任务的 *mAP* 均值的结果.

从表 2 中, 我们有以下发现:

1)SMCR 的表现优于所有前沿算法, 包括基于统计关联分析的方法和基于深度学习的方法. 其中 SMCR 方法在前 5 个、前 25 个、前 50 个的检索结果上的 *mAP* 均值在 2 个数据集上均优于目前最前沿的 SSACR 算法. 这表明, 虽然 SSACR 同样建模了媒体内语义损失和媒体间语义损失, SMCR 引入语义内的媒体约束模块, 通过更加逼真地映射出难以区分媒体的特征表示, 有助于进一步提升跨媒体检索性能.

Table 2 Comparison of Cross-Media Retrieval Performance on SciTechDaily and Wikipedia Datasets

表 2 在 SciTechDaily 和 Wikipedia 数据集上的跨媒体检索性能比较

数据集	算法	*mAP*@5			*mAP*@25			*mAP*@50		
		txt2img	img2txt	均值	txt2img	img2txt	均值	txt2img	img2txt	均值
SciTechDaily	CCA	0.233 7	0.180 6	0.207 1	0.232 8	0.176 1	0.204 4	0.222 5	0.178 9	0.200 7
	JFSSL	0.398 4	0.285 2	0.341 8	0.381 7	0.277 7	0.329 7	0.369 9	0.264 7	0.317 3
	CMDN	0.448 3	0.351 4	0.399 8	0.429 9	0.344 3	0.387 1	0.420 6	0.322 9	0.371 7
	ACMR	0.513 1	0.438 2	0.475 6	0.494 3	0.447 1	0.470 7	0.496 6	0.425 9	0.461 2
	DSCMR	0.504 2	0.457 7	0.480 9	0.481 2	0.464 6	0.472 9	0.481 0	**0.446 7**	0.463 8
	SSACR	0.509 1	0.457 2	0.483 1	0.504 9	0.448 7	0.476 8	0.507 2	0.435 5	0.471 3
	SMCR(本文)	**0.527 0**	**0.479 0**	**0.503 0**	**0.529 1**	**0.472 7**	**0.500 9**	**0.519 1**	0.442 6	**0.480 8**
Wikipedia	CCA	0.263 9	0.215 4	0.239 6	0.288 3	0.225 5	0.256 9	0.257 5	0.215 2	0.236 3
	JFSSL	0.443 2	0.348 1	0.395 6	0.426 6	0.352 8	0.389 7	0.415 2	0.347 9	0.381 5
	CMDN	0.526 5	0.419 4	0.472 9	0.504 6	0.417 1	0.460 8	0.487 4	0.393 8	0.440 6
	ACMR	0.637 2	0.492 0	0.564 6	0.625 1	0.493 7	0.559 4	0.588 7	0.482 4	0.535 5
	DSCMR	0.641 3	0.496 3	0.568 8	0.651 4	0.508 2	0.579 8	0.645 2	**0.497 3**	0.571 2
	SSACR	0.664 2	0.492 7	0.578 4	0.660 8	**0.508 9**	0.584 8	0.641 6	0.495 6	0.568 6
	SMCR(本文)	**0.701 4**	**0.505 9**	**0.603 6**	**0.671 4**	0.500 3	**0.585 8**	**0.650 3**	0.495 9	**0.573 1**

注: 黑体数值表示最优值.

2）SMCR 和JFSSL,CMDN,ACMR,DSCMR,SSACR 等同时建模媒体内相似性和媒体间相似性的模型，效果优于基于图像-文本对建模媒体间相似性的 CCA，表明同时考虑媒体内相似性和媒体间相似性能够提高跨媒体检索精度.

3）SMCR 和 ACMR, DSCMR, SSACR 的跨媒体检索性能优于在多任务学习框架中同样建模了媒体间不变性和媒体内判别性的 CMDN，表明对抗学习有助于进一步提升媒体间不变性和媒体内判别性的建模.

4）SMCR 通过分别建模相同语义、不同媒体数据在映射前和映射后的语义相似性，表现优于仅建模相同语义、不同媒体间数据在映射后的语义相似性的 ACMR 和 DSCMR. 这表示建模不同媒体的数据在映射前后的语义不变性有助于提高跨媒体检索精度.

5）SMCR 和所有前沿算法在 SciTechDaily, Wikipedia 这 2 个数据集上的表现一致，表明 SMCR 算法不仅局限于跨媒体科技资讯的检索，而且在通用的跨媒体检索任务中同样具备良好效果.

5.2 SMCR 方法主要组成部分的贡献

为了回答研究问题 2，我们将 SMCR 与去掉媒体间语义损失L_{imi}的 SMCR、去掉语义内媒体损失L_{con}的 SMCR 在 SciTechDaily 和 Wikipedia 这 2 个数据集上进行对比. 由于采用标签分类建模的媒体内语义损失L_{imd}并非本文创新，因此不对去掉L_{imd}的 SMCR 进行对比，结果如表 3、表 4 所示. 从表 3、表 4 中有 2 点发现：

1）去掉媒体间语义损失L_{imi}的 SMCR 和去掉语义内媒体损失L_{con}的 SMCR，相比 SMCR，跨媒体检索 *mAP* 值均有所下降. 这表明在特征映射网络中同时优化媒体间语义损失L_{imi}和语义内媒体损失L_{con}相比单独优化其中一个更有助于提升跨媒体检索表现.

2）SMCR 与其变体在 SciTechDaily, Wikipedia 这 2 个数据集上的跨媒体检索表现一致，再次表明 SMCR 方法并不局限于跨媒体科技资讯检索，而在通用的跨媒体检索任务上同样有效.

Table 3 Performance of SMCR and Its Variants in SciTechDaily Dataset

表 3 SMCR 与其变种在 SciTechDaily 数据集上的表现

本文方法	*mAP*	txt2img	img2txt	均值
SMCR（去掉L_{imi}）	*mAP*@5	0.519 6	0.462 7	0.491 1
	mAP@25	0.518 7	0.452 5	0.485 6
	mAP@50	0.502 4	0.440 8	0.471 6
SMCR（去掉L_{con}）	*mAP*@5	0.515 5	0.451 3	0.483 4
	mAP@25	0.507 3	0.447 4	0.477 3
	mAP@50	0.497 2	0.438 6	0.467 9
SMCR	*mAP*@5	0.527 0	0.479 0	0.503 0
	mAP@25	0.529 1	0.472 7	0.500 9
	mAP@50	0.519 1	0.442 6	0.480 8

Table 4 Performance of SMCR and Its Variants in Wikipedia Dataset

表 4 SMCR 与其变体在 Wikipedia 数据集上的表现

本文算法	*mAP*	txt2img	img2txt	均值
SMCR（去掉L_{imi}）	*mAP*@5	0.691 9	0.498 3	0.595 1
	mAP@25	0.662 2	0.493 7	0.577 9
	mAP@50	0.641 8	0.490 1	0.565 9
SMCR（去掉L_{con}）	*mAP*@5	0.680 6	0.503 8	0.592 2
	mAP@25	0.659 6	0.498 0	0.578 8
	mAP@50	0.641 6	0.493 8	0.567 7
SMCR	*mAP*@5	0.701 4	0.505 9	0.603 6
	mAP@25	0.671 4	0.500 3	0.585 8
	mAP@50	0.650 3	0.495 9	0.573 1

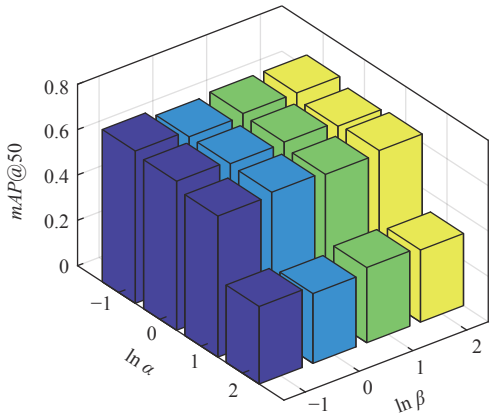
5.3 SMCR 方法的参数敏感性

本节回答研究问题 3. 式(8)中的特征映射网络的映射性损失L_{emb}有α和β这 2 个参数，分别控制媒体间语义损失L_{imi}和语义内媒体损失L_{con}在整体映射性损失L_{emb}中的参与度. 本节在 Wikipedia 数据集上改变α和β的取值，以测试 SMCR 算法的参数敏感性. 将α和β分别取值 0.1, 1, 10, 100, 特别而言，当$\alpha=0$时 SMCR 退化为去掉媒体间语义损失L_{imi}的 SMCR；当$\beta=0$时 SMCR 退化为去掉语义内媒体损失L_{con}的 SMCR. 因此α和β的取值不为 0. 固定一个参数(如α)的前提下，改变另一个参数(如β)进行实验，并采用 *mAP*@50 分别评估文本检索图像效果、图像检索文本效果、平均检索效果，结果如图 2 所示.

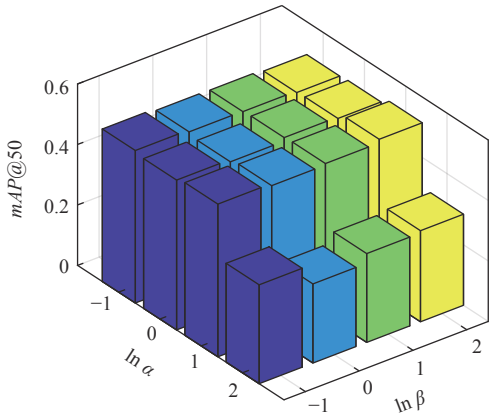
从图 2 中可见，当α取值为 0.1, 1, 10 和β取值为 0.1, 1, 10, 100 时，SMCR 表现较好. 这表明 SMCR 对参数不敏感，即泛化能力较好. 特别地，在文本检索图像任务上，当$\alpha=0.1$且$\beta=0.1$时，SMCR 表现最优；在图像检索文本任务上，当$\alpha=1$且$\beta=-1$时，SMCR 取得最优检索效果；在平均检索效果上，当$\alpha=-1$且$\beta=-1$时，SMCR 表现最好.

6 结 论

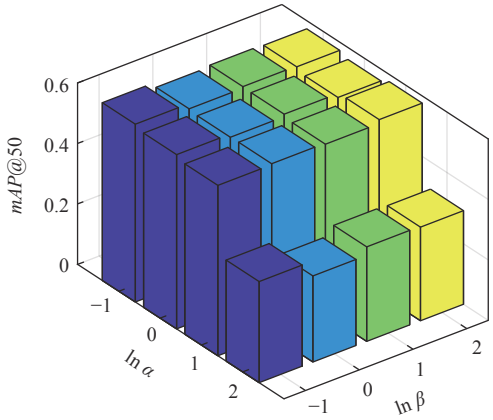
本文提出一种面向科技资讯的基于语义对抗和



(a) 文本检索图像效果



(b) 图像检索文本效果



(c) 平均检索效果

Fig. 2 Retrieval performance with α and β in Wikipedia dataset

图 2 Wikepedia 数据集上在α和β下的检索效果

媒体对抗的跨媒体检索方法(SMCR),能够同时学习跨媒体检索中的媒体内判别性、媒体间一致性、语义内判别性表示. SMCR 基于对抗学习方法,在极小化极大化游戏中涉及 2 个过程:生成具有媒体内判别性、媒体间一致性、语义间判别性表示的特征映射网络和试图辨别给定数据原始媒体的媒体判别网络. 本文引入媒体间一致性损失,以确保映射前后的媒体间数据保留语义一致性;此外,引入语义内媒体判别性损失,以确保映射后的数据在语义上接近自身,媒体上远离自身来增强特征映射网络混淆媒体判别网络的能力. 在 2 个跨媒体数据集上进行的综合实验结果证明了 SMCR 方法的有效性,且在跨媒体检索上的表现优于最前沿的方法.

作者贡献声明:李昂负责论文初稿撰写及修改、实验设计验证与核实;杜军平负责论文审阅与修订、研究课题监管与指导;寇菲菲负责指导实验方法设计;薛哲负责指导论文选题;徐欣和许明英负责实际调查研究;姜阳负责数据分析与管理.

参 考 文 献

[1] Fang Binxing. Building a new era of cyberspace security, promoting new trends in industry university research cooperation[J]. Science & Technology Industry of China, 2022, 36(2): 4−5 (in Chinese)
(方滨兴. 建设新时代网络空间安全产学研合作促进新态势[J]. 中国科技产业, 2022, 36(2): 4−5)

[2] Peng Yuxin, Huang Xin, Zhao Yunzhen. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(9): 2372−2385

[3] Kaur P, Pannu S, Malhi K. Comparative analysis on cross-modal information retrieval: A review [J]. Computer Science Review, 2021, 39(1): 100336

[4] Feng Xia, Hu Zhiyi, Liu Caihua. Survey of research progress on cross-modal retrieval[J]. Computer Science, 2021, 48(8): 13−23 (in Chinese)
(冯霞, 胡志毅, 刘才华. 跨模态检索研究进展综述[J]. 计算机科学, 2021, 48(8): 13−23)

[5] Liu Xingbo, Nie Xiushan, Yin Yilong. Mutual linear regression based supervised discrete cross-modal hashing[J]. Journal of Computer Research and Development, 2020, 57(8): 1707−1714 (in Chinese)
(刘兴波, 聂秀山, 尹义龙. 基于双向线性回归的监督离散跨模态散列方法[J]. 计算机研究与发展, 2020, 57(8): 1707−1714)

[6] Li Weiling, Tang Yong, Chen Guohua, et al. Implementation of academic news recommendation system based on user profile and message semantics [C] //Proc of the 9th Int Symp on Intelligence Computation and Applications. Berlin: Springer, 2017: 531−540

[7] Salehi S, Du J, Ashman H. Examining personalization in academic web search [C] //Proc of the 26th ACM Conf on Hypertext & Social Media (HT'15). New York: ACM, 2015: 103−111

[8] Wei Yunchao, Zhao Yao, Zhu Zhenfeng, et al. Modality-dependent cross-media retrieval[J]. ACM Transactions on Intelligent Systems and Technology, 2016, 7(4): 1−13

[9] Zhang Lu, Cao Feng, Liang Xinyan, et al. Cross-modal retrieval with correlation feature propagation[J]. Journal of Computer Research and

Development, 2022, 59(9): 1993−2002 (in Chinese)
(张璐, 曹峰, 梁新彦, 等. 基于关联特征传播的跨模态检索[J]. 计算机研究与发展, 2022, 59(9): 1993−2002)

[10] Wang Kaiye, He Ran, Wang Wei, et al. Learning coupled feature spaces for cross-modal matching [C] //Proc of the 13th IEEE Int Conf on Computer Vision (ICCV'13). Piscataway, NJ: IEEE, 2013: 2088−2095

[11] Hu Weiming, Gao Jun, Li Bing, et al. Anomaly detection using local kernel density estimation and context-based regression[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 32(2): 218−233

[12] Hardoon D, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: An overview with application to learning methods[J]. Neural Computation, 2004, 16(12): 2639−2664

[13] 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, 2011, 38(10): 2010−2023

[14] Zhai Xiaohua, Peng Yuxin, Xiao Jianguo. Learning cross-media joint representation with sparse and semisupervised regularization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(6): 965−978

[15] 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

[16] Feng Fangxiang, Wang Xiaojie, Li Ruifan. Cross-modal retrieval with correspondence autoencoder [C] //Proc of the 22nd ACM Int Conf on Multimedia (MM'14). New York: ACM, 2014: 7−16

[17] Yan Fei, Mikolajczyk K. Deep correlation for matching images and text [C] //Proc of the 33rd IEEE Conf on Computer Vision and Pattern Recognition (CVPR'15). Piscataway, NJ: IEEE, 2015: 3441−3450

[18] Peng Yuxin, Huang Xin, Qi Jinwei. Cross-media shared representation by hierarchical learning with multiple deep networks [C] //Proc of the 25th Int Joint Conf on Artificial Intelligence (IJCAI'16). Palo Alto, CA: AAAI, 2016: 3846−3853

[19] Kou Feifei, Du Junping, He Yijiang, et al. Social network search based on semantic analysis and learning[J]. CAAI Transactions on Intelligence Technology, 2016, 1(4): 293−302

[20] Xu Liang, Du Junping, Li Qingping. Image fusion based on nonsubsampled contourlet transform and saliency-motivated pulse coupled neural networks [J]. Mathematical Problems in Engineering, 2013, 19(1): 135182

[21] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning [C] //Proc of the 28th Int Conf on Machine Learning (ICML'11). New York: ACM, 2011: 689−696

[22] He Li, Xu Xing, Lu Huimin, et al. Unsupervised cross-modal retrieval through adversarial learning [C] //Proc of the 18th IEEE Int Conf on Multimedia and Expo (ICME'17). Piscataway, NJ: IEEE, 2017: 1153−1158

[23] Li Chao, Deng Cheng, Li Ning, et al. Self-supervised adversarial hashing networks for cross-modal retrieval [C] //Proc of the 36th IEEE Conf on Computer Vision and Pattern Recognition (CVPR'18). Piscataway, NJ: IEEE, 2018: 4242−4251

[24] Wang Bokun, Yang Yang, Xu Xing, et al. Adversarial cross-modal retrieval [C] //Proc of the 25th ACM Int Conf on Multimedia (MM'17). New York: ACM, 2017: 154−162

[25] Zhen Liangli, Hu Peng, Wang Xu, et al. Deep supervised cross-modal retrieval [C] //Proc of the 37th IEEE Conf on Computer Vision and Pattern Recognition (CVPR'19). Piscataway, NJ: IEEE, 2019: 10386−10395

[26] Liu Chong, Du Junping, Zhou Nan. A cross media search method for social networks based on adversarial learning and semantic similarity[J]. SCIENTIA SINICA Informationis, 2021, 51(5): 779−794 (in Chinese)
(刘翀, 杜军平, 周南. 一种基于对抗学习和语义相似度的社交网络跨媒体搜索方法[J]. 中国科学: 信息科学, 2021, 51(5): 779−794)

[27] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C] //Proc of the 27th Int Conf on Neural Information Processing Systems (NIPS'14). Cambridge, MA: MIT Press, 2014: 2672−2680

[28] Li Chao, Deng Cheng, Li Ning, et al. Self-supervised adversarial hashing networks for cross-modal retrieval [C] //Proc of the 36th IEEE Conf on Computer Vision and Pattern Recognition (CVPR'18). Piscataway, NJ: IEEE, 2018: 4242−4251

[29] Yu Chaohui, Wang Jindong, Chen Yiqiang, et al. Transfer learning with dynamic adversarial adaptation network [C] //Proc of the 19th IEEE Int Conf on Data Mining (ICDM'19). Piscataway, NJ: IEEE, 2019: 778−786

[30] Xue Zhe, Du Junping, Du Dawei, et al. Deep low-rank subspace ensemble for multi-view clustering[J]. Information Sciences, 2019, 482(5): 210−227

[31] Fang Yuke, Deng Weihong, Du Junping, et al. Identity-aware CycleGAN for face photo-sketch synthesis and recognition [J]. Pattern Recognition, 2020, 102(6): 107249

[32] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks [J]. arXiv preprint, arXiv: 1511.06434, 2015

[33] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation [C] //Proc of the 32nd Int Conf on Machine Learning (ICML '15). New York: ACM, 2015: 1180−1189

[34] Hoffer E, Ailon N. Deep metric learning using triplet network [C] //Proc of the 3rd Int Workshop on Similarity-Based Pattern Recognition. Berlin: Springer, 2015: 84−92

[35] Liang Xiaodan, Zhang Hao, Lin Liang, et al. Generative semantic manipulation with mask-contrasting GAN [C] //Proc of the 15th European Conf on Computer Vision (ECCV'18). Berlin: Springer, 2018: 574−590

[36] Xiong Wei, Luo Wenhan, Ma Lin, et al. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks [C] //Proc of the 36th IEEE Conf on Computer Vision and Pattern Recognition (CVPR'18). Piscataway, NJ: IEEE, 2018: 2364−2373

[37] SCITECHDAILY. SciTechDaily [EB/OL]. [2022-01-01].https://scitechdaily.com/news/technology

[38] Costa J, 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

[39] Peng Yuxin, Huang Xin, Qi Jinwei. Cross-media shared representation by hierarchical learning with multiple deep networks [C] //Proc of the 25th Int Joint Conf on Artificial Intelligence (IJCAI'16). Palo Alto, CA: AAAI, 2016: 3846−3853

[40] Ranjan V, Rasiwasia N, Jawahar C. Multi-label cross-modal retrieval [C] //Proc of the 15th IEEE Int Conf on Computer Vision (ICCV'15). Piscataway, NJ: IEEE, 2015: 4094−4102

Li Ang, born in 1993. PhD candidate. Member of CCF. His main research interests include information retrieval, data mining, and machine learning.

李　昂, 1993 年生. 博士研究生. CCF 会员. 主要研究方向为信息检索、数据挖掘、机器学习.

Du Junping, born in 1963. Professor. Fellow of CCF. Her main research interests include artificial intelligence, machine learning, and pattern recognition.

杜军平, 1963 年生. 教授. CCF 会士. 主要研究方向为人工智能、机器学习、模式识别.

Kou Feifei, born in 1989. Lecturer. Member of CCF. Her main research interests include semantic learning and multimedia information processing.

寇菲菲, 1989 年生. 讲师. CCF 会员. 主要研究方向为语义学习和多媒体信息处理.

Xue Zhe, born in 1987. Associate professor. Member of CCF. His main research interests include machine learning, artificial intelligence, data mining, and image processing.

薛　哲, 1987 年生. 副教授. CCF 会员. 主要研究方向为机器学习、人工智能、数据挖掘、图像处理.

Xu Xin, born in 1992. PhD. Member of CCF. Her main interests include knowledge graph, information retrieval, and machine learning.

徐　欣, 1992 年生. 博士. CCF 会员. 主要研究方向为知识图谱、信息检索、机器学习.

Xu Mingying, born in 1987. PhD. Member of CCF. Her main research interests include intelligent information retrieval, science & technology big data analysis, and data mining.

许明英, 1987 年生. 博士. CCF 会员. 主要研究方向为智能信息检索、科技大数据分析、数据挖掘.

Jiang Yang, born in 1995. Master. His main research interests include nature language processing, cross-media retrieval, and deep learning.

姜　阳, 1995 年生. 硕士. 主要研究方向为自然语言处理、跨媒体检索、深度学习.