ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 3,May 2023,pp. 479－505
DOI:10. 16337/j. 1004-9037. 2023. 03. 001

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

图文跨模态检索研究进展

张飞飞[1]，马泽伟[1]，周　玲[2]，孟铃涛[1]

(1. 天津理工大学计算机科学与工程学院，天津 300384;2. 中南大学交通运输工程学院，长沙 410083)

摘　要：随着互联网技术的迅速发展,文本和图像等各种类型的数据在网络上呈现爆发式增长,如何从这些多源异构且语义关联的多模态数据中获取有价值的信息则尤为重要。跨模态检索能够突破模态的限制,跨越不同模态的数据进行信息检索,满足用户获取有关事件信息的需求。近年来,跨模态检索已经成为了学术界和工业界研究的热点问题。本文聚焦于图文跨模态检索任务,首先介绍图文跨模态检索的定义,并分析说明了当前该任务面临的挑战。其次,对现有的研究方法进行归纳总结,将其分为3大类:(1)传统方法;(2)基于深度学习的方法;(3)基于哈希表示的方法。然后,详细介绍了图文跨模态检索的常用数据集,并对常用数据集上已有算法进行详细分析与比较。最后,对图文跨模态检索任务的未来发展方向进行展望。

关键词：多模态学习;图文跨模态检索;深度学习;自监督学习;哈希学习

中图分类号：TP391.4　　**文献标志码**：A

Recent Advances in Cross Modal Image Text Retrieval

ZHANG Feifei[1], MA Zewei[1], ZHOU Ling[2], MENG Lingtao[1]

(1. School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China;2. School of Traffic & Transportation Engineering, Central South University, Changsha 410083, China)

Abstract: With the rapid development of Internet technology, the volume of different types of data has grown tremendously, such as texts and images. How to obtain valuable information from such heterogeneous but semantic related multimodal data is particularly important. Cross-modal retrieval is an essential way to meet users' requirements for obtaining different information on the Internet, which can effectively deal with the multimodal data. In recent years, cross modal retrieval has become a hot issue in both academic and industrial area. In this paper, we make a comprehensive overview of the image-text cross modal retrieval task, including definitions, challenges, and detailed discussions about the existing methods. Specifically, we first divide the existing methods into three main categories: (1) traditional methods, (2) methods based on deep learning; and (3) Hash based representation method. Then, we introduce the commonly used cross-modal retrieval benchmarks and discuss the existing methods on these benchmarks in detail. Finally, the future development direction of image-text cross modal retrieval task is prospected.

Key words: multimodal learning; image-text retrieval; deep learning; self-supervised learning; Hash learning

基金项目:国家重点研发计划(2018AAA0102200);国家自然科学基金(62036012,62002355,62072455,62102415,62106262);天津市自然科学基金(22JCYBJC00030)。
收稿日期:2022-12-15;**修订日期**:2023-02-20

引　言

随着互联网的发展，每时每刻都有巨量的文本、图像和视频被上传到互联网上，这些不同类型的数据可以用来描述同一事件或者话题，但是它们之间的信息分布差异很大。例如，图像是外界存在的实体场景信息，而文本是由人类自己编写的离散信息，这种不同类型的数据被称为多模态数据。为了对网络中的大量数据进行有效组织和管理，不同类型的检索方法被研究人员提出[1-5]，例如文本检索[1]、图像检索[2]以及视频检索[3]等。这些单模态检索方法虽然能够取得较好的检索结果，但是只能进行单一模态数据之间的检索。在多模态信息时代，人们不再满足于单一模态知识的获取，而是希望获得多种模态的搜索结果。比如，当用户学习制作美食时，可以使用美食的名称或者照片检索出对应制作方法的文本描述[6]，不同模态的检索结果可以相互补充，使用户更全面地理解检索到的信息。跨模态检索旨在用一种模态的数据查询与之语义相关的另一模态数据，例如以图像检索文本[7]和以文本检索视频[8]等。其中，图像文本跨模态检索是一项基础性任务，其任务成果能够拓宽其他模态检索任务的研究思路。所以对图像文本跨模态检索的研究是多模态领域的一个热点方向，具有十分重要的研究意义。为了便于研究人员更好地了解图像文本跨模态检索（以下简称为跨模态检索），本文将对该任务的定义、挑战、相关数据集，以及相关算法进行全面分析与讨论。

跨模态检索主要包括两个子任务，即用图像搜索文本和用文本搜索图像，如图1所示。该任务的关键挑战是语义鸿沟问题，即图像和文本的底层表示不一致，分布在不同特征空间中，很难对其进行有效的相似性度量。研究人员通过分析多模态数据的语义关联性，提出许多解决方法[9-10]。例如，将图像和文本的特征投影到一个公共特征空间中，利用数据的标签信息或成对信息学习衡量图像和文本间的相似度。

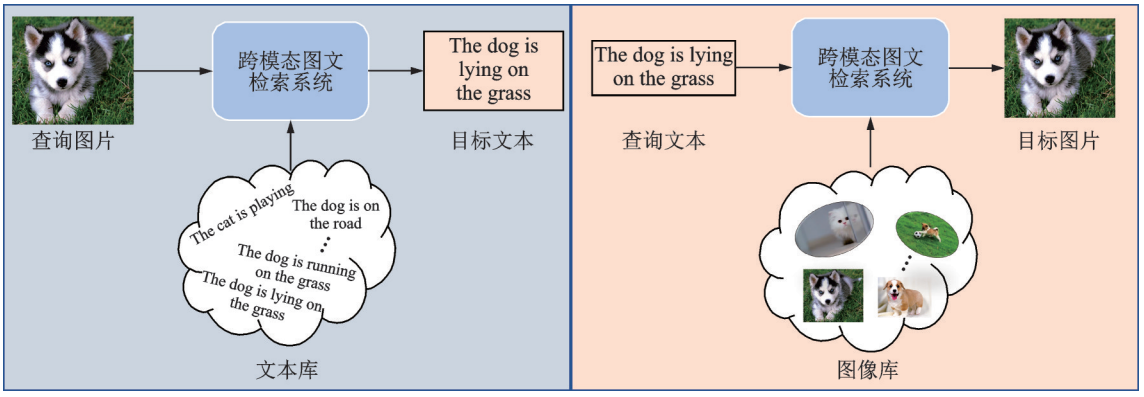


图1　跨模态检索任务定义

Fig.1　Cross modal retrieval task definition

虽然跨模态检索已经得到研究人员的广泛关注，且已有大量相关工作产出，但上述挑战仍未完全得到解决，检索性能也有待提高。对原来的工作进行总结整理和对比分析，有助于在原有工作的基础上找到更好的解决方案。目前也已有很多优秀的综述对跨模态检索任务的相关方法进行归纳总结，例如Liu等[11]对跨模态检索传统方法进行了详细介绍和梳理总结，总结的方法涵盖了多模态特征表示、多模态语义理解和多模态相关性研究3大类，同时介绍了每种方法的动机、细节及其缺点。Xu等[12]对多视图学习展开综述，重点从协同训练、多核学习以及公共子空间学习的角度进行梳理，然而此综述并非针对异构数据的检索任务。Wang等[13]从特征表示的角度出发，通过实值表示学习和二进制表示学习

对跨模态检索展开详细阐述,同时提出跨模态检索的未来发展方向,包括收集大规模数据集、合理利用噪声样本、进一步研究深度学习方法以及细粒度关系建模。Peng等[14]从语义鸿沟这一难点出发,通过基于典型相关性分析、基于神经网络、基于图正则化方法、基于度量学习的方法以及基于跨模态哈希方法对跨模态检索任务进行了介绍。

尽管以上综述从不同角度对跨模态检索任务进行了较全面的总结,但仍存在以下不足:(1)它们都侧重于对跨模态检索现有的方法进行总结和分类,没有考虑图文跨模态检索是特殊的跨模态检索子任务,因此没有专门对图文跨模态检索展开讨论,对图像和文本两种模态的针对性不强;(2)监督信息是影响跨模态检索性能的重要因素,而文献[11,12,14]没有从监督信息的角度对跨模态检索进行阐述,文献[13]虽然按照标签类型进行了分类,但是忽视了基于预训练任务的自监督学习方法;(3)在介绍各数据集方法对比部分,以上综述仅仅以表格的形式列出了每种方法的性能,未对每种对比方法进行介绍,也未对其进行详细对比分析。因此,本文将从图像-文本的角度对跨模态检索任务进行全面的总结与分析。

1 图文跨模态检索挑战

由于不同模态特征的差异,使得跨模态检索面临诸多挑战。现有的跨模态检索模型大多是基于构造公共特征空间的框架,先提取图像特征和文本特征,再将两种特征映射到公共空间中进行相似性比较。在此训练过程中为了更好地提升模型的性能,研究人员需要考虑如何跨越语义鸿沟、如何去除冗余特征以及如何提升检索效率和精度等问题。

1.1 如何跨越语义鸿沟

跨模态检索使用的数据具有“底层特征异构、高层语义相关”的特点,比如描述狗躺在地上的文本和一张狗躺在草地上的图像在语义上是相似的,但是其特征表示却千差万别。文本通常展现的是一种离散的高层语义信息,而图像则由连续的底层像素特征来表示,这导致不同模态间具有低级特征和高级概念的差异,被称为模态间的语义鸿沟,使得衡量不同模态间相似性变得非常困难,因此同一语义的不同模态数据特征之间的特征异构是跨模态检索的重点问题。

1.2 如何去除冗余特征

信息论中定义描述息的数据是由有用信息和无用信息构成。有用信息是研究人员最终希望提取的部分特征信息,无用信息就是与上下文语境无关的部分,比如一张狗的图像对应的有用信息就是图像中狗所在区域的特征。如何从一张图像中提取出与文本实体相对应的区域特征是一件困难的事,而且文本信息相较于图像信息更为抽象,包括动词、名词以及无实质含义的其他词汇,语义更复杂,有用信息更多。因此,研究人员常使用注意力机制[15-18]和构建场景图[19]等方法去除冗余特征保留有用特征。如何选取不同模态间相对应的特征值来进行相似性度量是跨模态检索的一大难点。

1.3 如何提升检索效率和精度

社交媒体上的多媒体数据每天都在不断增长,需要高效的信息检索来支撑[20]。大多数现有模型通过不断加大模型规模[21]或者增加模型层数[22]的方式追求检索精度。然而,在此过程中增加了模型的训练参数,导致模型在训练的时候需要耗费大量的算力和时间,并且在检索时也存在检索速度缓慢的问题[23],使得模型无法在现实中应用。研究人员通常使用跨模态哈希相关算法[24-26]解决该问题,但此类方法由于哈希码的表达能力有限,导致模型无法取得很好的性能。因此如何保证在提升模型精度的同时提升模型的效率,是后续研究有待解决的棘手问题。

2 图文跨模态检索相关研究

跨模态检索方法分类架构如图2所示。给定一个图像文本对，通过视觉编码器和文本编码器分别对其进行特征提取。为了使模型学习到正确的图文匹配关系，现有的研究方法主要包括以下3方面。(1)传统方法利用手工特征对不同数据进行特征表示，然后通过优化统计值来学习子空间的投影矩阵。根据方法侧重点的不同，将传统方法分为特征提取方法和相似性度量方法。(2)基于深度学习的方法利用神经网络非线性投影特性得到语义表达能力更强的语义特征。本文将深度学习方法分为无监督方法、有监督方法以及自监督方法。无监督方法不需要借助任何标签信息，通过在特征空间中最大化同一样本不同模态表征的相似性来学习跨模态数据之间的关联；有监督方法利用手工标注的标签信息(比如COCO数据集[27]中每张图片都有对应的标签向量来储存该图片中存在的物体种类)学习多模态数据的语义表示；自监督方法利用辅助任务从不同模态数据中挖掘数据本身的监督信息。(3)基于哈希表示的方法通过哈希映射学习一个统一的二进制空间。由于目前针对自监督哈希跨模态的研究较少，所以本文不总结该类方法，按照是否使用监督信息将基于哈希表示的方法分为监督学习方法和无监督学习方法。下面将围绕以上3个类别对已有方法进行详细分析与介绍。

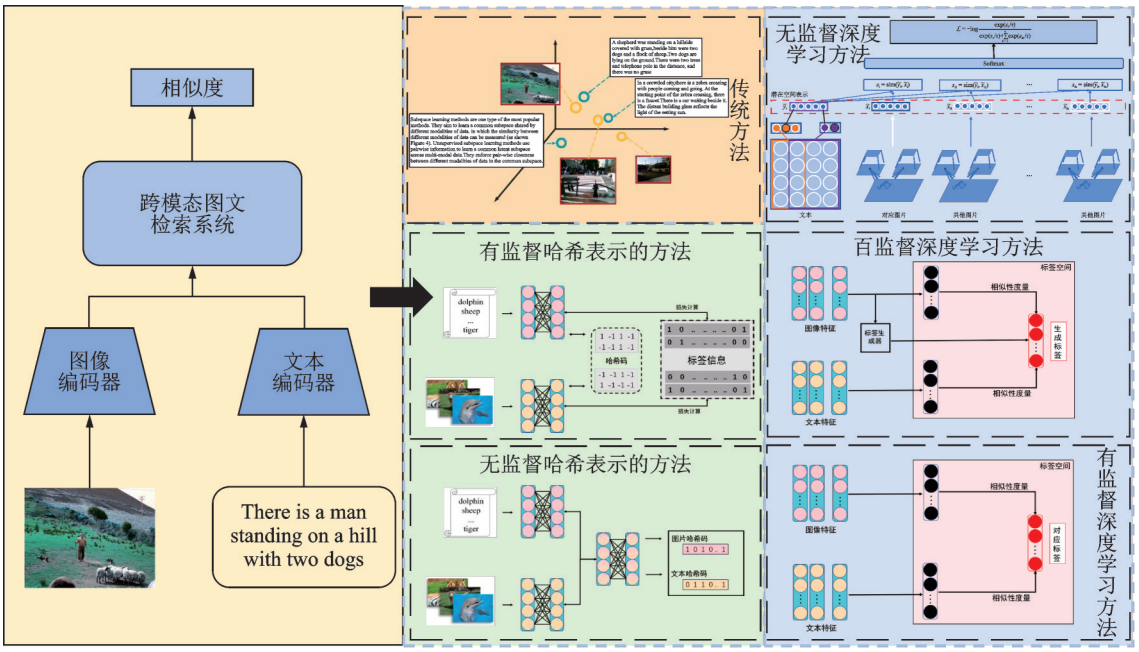


图2 跨模态检索方法的分类架构

Fig.2 Classification architecture for cross modal retrieval methods

2.1 传统方法

传统方法主要通过传统手工特征对不同数据进行特征表示，文本特征如BoW[28]、LDA[29]等，图像特征如SIFT[30]、HOG[31]等。然后通过构造公共表示空间进行相似性度量。本节将分为特征提取和相似性度量两个部分对传统方法进行详细介绍，其中具有代表性的方法如表1所示。

2.1.1 特征提取

传统图文跨模态检索方法使用文本和图像两种模态的手工特征作为输入。其中文本特征大多使用基于统计分析方法处理的向量特征。例如，Kolenda等[37]提出的多模态内容度量检索方法中就通过

表1 代表性传统方法简要介绍

Table 1 Brief introduction to representative traditional methods

时间	作者	代表方法	优势	局限性
2004年	Hardoon等	CCA[32]	模型构造简单,能够很好地处理不同模态间的线性关系。	仅能同时处理两种模态信息,无法对非线性关系进行建模。
2006年	Zhen等	KCCA[33]	通过使用核函数使得模型能够处理非线性关系。	方法的计算复杂度高,容易过拟合,核函数的选择也比较困难。
2007年	Kim等	TCCA[34]	提取不同模态数据的高阶协方差张量作为特征矩阵,使得模型能够同时处理多个模态数据。	需要计算高维协方差矩阵,导致模型效率低下。
2014年	Rasiwasia等	Cluster-CCA[35]	能够利用标签信息监督模型训练,同时利用核函数来处理非线性关系。	需要手工标注的标签信息,并且模型计算量大,效率低。
2020年	Xu等	P3S[36]	利用共享子空间和私有子空间之间的正交性约束来排除不重要信息,提升模型对噪声的鲁棒性。	该方法在排除不重要信息的同时容易收到数据集偏差的影响,导致模型对一些重要的场景信息不敏感。

词袋法(Bag of words, BoW)来提取文本特征。词袋法是早期使用最多的文本信息提取方法[28],该方法将文本中的单词进行汇总形成词袋,并组合成一个向量来表示该文本,向量的大小为词袋中包含的不重复的单词数量,向量中每一位的值代表对应的单词是否在文本中出现。但是词袋法的缺点也很明显,由于使用字典向量表示一个文本会导致向量中单词是无序的,因此忽略了文本信息中顺序和语法的信息。由于BoW不能反映出单词之间的关系,BLei等[29]提出了潜在狄利克雷(Latent Dirichlet allocation, LDA)模型,该模型能够获取输入文本中的隐含主题和概念结构信息,因此这种方法得到了早期研究人员的广泛使用[38-40]。关于图像的传统表示,有颜色、纹理、形状和梯度等底层特征。虽然这些特征可以一定程度上反映图像的像素分布,但却无法区分出前景和背景以及图像中的旋转和遮挡等情况。随后,研究人员提出了具有不变性的局部特征[30-31],并且在跨模态检索领域得到了广泛的应用。例如,Lowe等[30]提出的(Scale invariant feature transform, SIFT)通过在空间中寻找极值点以获得其大小比例、空间位置和旋转因子来构造图像的表达特征。在此基础上,Rasiwasia等[37]在提出的跨模态检索方法中使用LDA和SIFT分别提取文本特征和图像特征,使得模型获得了更好的性能。方向梯度直方图(Histogram of oriented gradient, HOG)[31]利用统计分析方法获得图像局部区域的梯度直方图来提取图像的特征信息,在Kang等[41]提出的跨模态一致性学习方法中被用作图像特征。

手工文本特征很难应用在大规模语料库上,因为随着语料库中词语的增多,文本向量的维度也会相应增大,而一个文档中出现的词语可能会远小于预料库的词语数,这就导致了文本向量是稀疏的,不但会浪费计算资源还会导致维度灾难。手工图像特征对图像尺寸以及简单的图像旋转具有鲁棒性,但是如果在图像出现部分丢失、光线变化及大尺度旋转的情况下会出现严重的误差,并且这种手工特征需要专业的领域知识,对于不同的图像或者应用需要设计不同的特征,因此效率较低。

2.1.2 相似性度量

传统方法通常使用统计分析方法来将不同模态特征映射到公共表示空间中,从而直接对模态特征进行相似性度量。Hardoon等[32]在2004年提出的典型相关性分析(Canonical correlation analysis, CCA)是其代表性方法,该方法将两组不同的变量映射到公共表示空间,然后使用矩阵分解等方法来最大化这两组变量的相关系数。假设存在文本和图像两种不同模态特征矩阵 $\boldsymbol{T}=\{t_1, t_2, t_3, \cdots, t_n\}$, $\boldsymbol{I}=$

$\{i_1, i_2, i_3, \cdots, i_n\}$,CCA对于这2个特征矩阵分别使用线性系数向量$\boldsymbol{\varphi}_t$和$\boldsymbol{\varphi}_i$来将其投影到公共表示空间中表示为$\boldsymbol{U}$和$\boldsymbol{V}$,然后构建$\boldsymbol{T}$和$\boldsymbol{I}$的模态内协方差矩阵$\boldsymbol{\Sigma}_{TT}$、$\boldsymbol{\Sigma}_{II}$和模态间协方差矩阵$\boldsymbol{\Sigma}_{TI}$,如式(1,2)所示。

$$\boldsymbol{U}=\boldsymbol{\varphi}_t^{\mathrm{T}}\boldsymbol{T},\quad \boldsymbol{V}=\boldsymbol{\varphi}_i^{\mathrm{T}}\boldsymbol{I} \tag{1}$$

$$\boldsymbol{\Sigma}_{TT}=\frac{1}{n}\sum_{q=1}^{n}\boldsymbol{t}_q\boldsymbol{t}_q^{\mathrm{T}},\quad \boldsymbol{\Sigma}_{II}=\frac{1}{n}\sum_{q=1}^{n}\boldsymbol{i}_q\boldsymbol{i}_q^{\mathrm{T}},\quad \boldsymbol{\Sigma}_{TI}=\frac{1}{n}\sum_{q=1}^{n}\boldsymbol{t}_q\boldsymbol{i}_q^{\mathrm{T}} \tag{2}$$

然后利用模态内协方差矩阵和模态间协方差矩阵来计算$\boldsymbol{U}$和$\boldsymbol{V}$的相似系数β,该系数反映了模态间特征的相关性,表达式为

$$\beta=\arg\max\frac{\boldsymbol{\varphi}_t^{\mathrm{T}}\boldsymbol{\Sigma}_{TI}\boldsymbol{\varphi}_i}{\sqrt{\boldsymbol{\varphi}_{tt}^{\mathrm{T}}\boldsymbol{\Sigma}_{TT}\boldsymbol{\varphi}_t}\sqrt{\boldsymbol{\varphi}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{II}\boldsymbol{\varphi}_i}} \tag{3}$$

在对式(3)计算之前需要对分母进行归一化处理来确定线性系数向量$\boldsymbol{\varphi}_t$和$\boldsymbol{\varphi}_i$,令$\boldsymbol{\varphi}_{tt}^{\mathrm{T}}\boldsymbol{\Sigma}_{TT}\boldsymbol{\varphi}_t=\boldsymbol{\varphi}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{II}\boldsymbol{\varphi}_i=1$,之后使用拉格朗日乘子法将问题转换为

$$\mathcal{L}(\boldsymbol{\varphi}_t,\boldsymbol{\varphi}_i)=\boldsymbol{\varphi}_t^{\mathrm{T}}\boldsymbol{\Sigma}_{TI}\boldsymbol{\varphi}_i-\frac{\lambda}{2}\left(\boldsymbol{\varphi}_{tt}^{\mathrm{T}}\boldsymbol{\Sigma}_{TT}\boldsymbol{\varphi}_t-1\right)-\frac{\theta}{2}\left(\boldsymbol{\varphi}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{II}\boldsymbol{\varphi}_i-1\right) \tag{4}$$

式中λ和θ为系数向量。最后对式(4)求偏导后得到最大化相关系数β的线性系数向量$\boldsymbol{\varphi}_t$和$\boldsymbol{\varphi}_i$,该思想为后续许多跨模态检索方法奠定了基础,但是其也存在一些缺点:(1)只能处理两种模态信息,不能计算3个或3个以上模态间数据的相关性;(2)只能计算向量之间的线性相关性,然而在大多数实际情况下,模态间的关系是非线性的;(3)仅考虑模态间相似性的衡量,却忽视了选取有效信息的重要性;(4)不能利用标签信息,缺失多模态类别语义的学习。

针对第1个问题,Tenenhaus等[42]提出正则广义典型相关性分析法(Regularized generalized canonical correlation analysis, RGCCA),该方法实现了对两种以上模态关系的约束。Carrol等[43]提出了方差最大化的广义典型相关性分析法,该方法将多个模态投影到一个公共特征空间中,在这个公共的子空间中最大化所有模态的相关性。另外,Kim等[34]提出了张量正则相关分析法(Tensor canonical correlation analysis, TCCA),该方法提取不同模态数据的高阶协方差张量作为特征,然后通过对多个模态数据的高维协方差矩阵进行分析,直接最大化所有模态数据的相关性。

针对第2个问题,Zheng等[33]提出核典型相关性分析法(Kernel canonical correlation analysis, KCCA),该方法使用核函数来处理非线性关系,首先使用核函数将数据映射到高维空间中,在此过程将线性关系转换为了非线性关系,最后在该空间中使用CCA进行求解。但是此方法的计算复杂度高,容易过拟合,核函数的选择也比较困难。为此,Lopez等[44]提出了随机非线性典型性相关分析法(Randomized nonlinear canonical correlation analysis, RCCA),RCCA通过对数据进行随机非线性映射来解决非线性相关问题,同时移除了核函数机制,在不影响模型精度的情况下降低了计算量。Sun等[45]通过保留数据的局部性来解决非线性问题,提出了局部保持典型相关性分析法(Proposed locality preserving canonical correlation analysis, LPCCA),该方法在保持特征信息的局部特征结构的同时对全局非线性特征进行降维,进而将非线性问题转化为线性问题进行求解。

针对第3个问题,Wang等[46]提出了联合特征子空间学习方法,该方法分别对不同模态的特征空间以及公共特征空间中的特征进行正则化处理,保留了多模态数据的判别性特征,从而保证模态内部和模态之间的语义相关性。Xu等[36]提出了私有共享子空间方法(Private-shared subspaces separation, P3S),利用共享子空间和私有子空间之间的正交性约束,使得该方法能排除图像中不相关的场景信息或者文本中错句产生的干扰,进一步去除了冗余信息。

针对第4个问题,Sun等[47]提出了判别式典型性分析法(Discriminative CCA, DisCCA),该方法在最小化类内相似性的同时最大化类间相似性,同时利用标签信息来监督自身训练。Rasiwasia等[38]提出了语义相关性匹配(Semantic correlation matching, SCM)方法,该方法对CCA做出了改进,通过定义语义概念词典获得标签信息,之后将文本特征和图像特征根据标签映射为后验概率向量,最大化对应概率向量的相似性。之后他们又提出了聚类相关性分析法(Cluster canonical correlation analysis,Cluster-CCA)[35],该方法利用标签信息将两组模态数据划分为多个集合,在特征空间上学习最大化2个集合之间相关性的判别性低维表示。类似于KCCA,Cluster-CCA通过引入核函数来将特征引入高维空间中来处理非线性问题,但是会导致计算量增大,模型效率变低。受文献[46]中正则化应用的启发,Xu等[48]提出了一种半监督语义保持跨模态检索方法(Semantic consistency cross-modal retrieval, SCCMR),该方法使用语义一致性正则化将数据表示转换为类别的概率分布,对齐标签空间中的不同模态特征。为了进一步克服模态间的语义鸿沟,提取不同模态更具代表性的语义特征,Wang等[49]提出了基于图正则化和模态依赖的跨模态检索方法(Graph regularization and modality dependence, GRMD),该方法利用原始特征的空间信息构造邻接图,并且利用类标签的语义信息缩小公共空间中不同模态间的语义差异,增强了不同模态间的潜在相关性,提升了模型的性能。

传统方法主要依靠统计分析,模型结构简单。但是缺点也很明显:模型通常以最大化公共空间特征相似性,对模态内数据细粒度信息提取和模态间语义对齐专注较少。并且由于模态间的关系往往是非线性的,传统方法对其效果不佳,因为需要储存所有的数据来计算不同数据间的协方差矩阵,因此在面对高维特征矩阵时,计算量也非常庞大。

2.2 基于深度学习的方法

深度学习模拟了人类大脑在对事物特征提取和分类方面的工作方式[50],具备很强的学习能力。基于深度学习的方法通过误差反向传播学习到更具非线性表达能力的特征表示,可以克服手工特征表示能力弱的缺点。下面将根据监督信息的使用方式,将基于深度学习的跨模态检索方法分为3类:无监督学习方法、有监督学习方法以及自监督学习方法。其中,无监督学习方法仅使用数据集中图像和文本的成对关系来学习不同模态数据的一一对应关系;有监督学习方法通常利用标签信息优化公共子空间,比如利用人工标注的图像中包含物体类别作为标签信息;自监督学习方法通过辅助任务从数据本身挖掘监督信息。本节将这3类方法进行详细介绍,在表2中列出了其中具有代表性的方法。同时为了直观地展示本节中不同类别方法之间的区别,从表2中选出基于无监督、有监督和自监督的3个模型进行结构展示。

表2 代表性基于深度学习方法简要介绍

Table 2 Brief introduction to representative deep learning methods

时间	作者	代表方法	优势	局限性
2017年	He等	UCAL[51]	利用生成对抗学习,将特征提取和模态信息识别转化为对抗训练,能够在提取特征的同时保持模态内信息不变。	在对抗训练的时候模型过度关注模态内部的特征,没有考虑模态之间的语义特征。
2018年	Zhen等	SCAN[16]	使用交叉注意力对不同模态的特征进行细粒度的语义对齐。	没有考虑到文本语义的多样性,而是平等地对齐所有模态信息。
2019年	Zhen等	DSCMR[52]	在训练过程中利用标签信息来监督模型训练,使用多个损失来进行语义对齐。	在进行模态间特征对齐的时候没有考虑保留原有模态特征的语义信息。
2020年	Li等	Oscar[22]	使用经过预训练的Transformer来构建模型,显著提高了模型的精度。	模型结构复杂,参数量巨大,训练效率低,难以轻量化部署。
2022年	Xu等	ELRCMR[53]	使用正则化和动态权重平衡策略提高了模型对于噪声样本的抵抗力。	模型在使用动态权重平衡策略来移动聚类中心,导致模态特征结构被破坏。

2.2.1 无监督学习方法

无监督深度学习方法通常将同一样本的不同模态特征投影到公共特征空间中,然后利用样本内部不同模态的对应关系来学习多模态数据的公共表示。下面将对基于无监督的深度跨模态检索方法进行详细介绍。

由于深度神经网络(Deep neural networks, DNN)可以很好地处理非线性关系,因此Andrew等[10]将典型相关分析CCA和深度神经网络结合起来,提出了深度典型性相关分析方法(Deep canonical correlation analysis, DCCA),该方法使用深度神经网络分别求出2个模态的非线性投影矩阵,然后最大化2个投影矩阵的相关性,解决了传统方法中使用核函数导致模型计算复杂的问题,并且提高了模型的性能。但是DCCA将研究重点放在了模态间相似性度量方面,并没有考虑模态内部的耦合特征选择问题。为此Feng等[54]提出基于通信自动编码器(Correspondence autoencoder,Corr-AE)的跨模态检索方法,该方法使用2个自编码器来对原始特征进行提取,通过单模态表征学习损失和多模态相似性损失来训练模型。其中单模态表征学习损失表示最小化每种模态的特征矩阵和原始特征矩阵的误差,保证编码器在重构每种模态内特征结构的同时保持原有的语义信息;多模态相似性损失最大化不同模态特征矩阵的相似性来学习模态间的共有信息。联合单模态表征学习损失和多模态相似性损失,使得模型同时考虑到了模态内的特征选择和模态间相关性度量,提升了模型的性能。以上方法大都依赖于低级的视觉特征和文本特征进行跨模态检索,没有考虑模态数据中包含的语义信息和空间信息。为了提取高级语义信息,Wang等[9]提出了模态特定特征的学习模型,该方法使用两种不同的神经网络来提取图像文本的公共空间表示。其中文本使用词嵌入学习网络来提取更高层次的文本语义特征,图像使用深度神经网络进行特征提取。最后使用极大似然方法来最小化同一样本不同模态的特征相似性距离和最大化不相关样本的相似性距离,进而提升模型语义对齐的能力。Zhao等[55]提出了同构和异构同现模型,其中同构表示统一模态,异构表示不同模态,该方法由同构同现和异构同现两个模块组成,其中同构同现模块从图像区域和单词的单一模态中捕获相邻节点的关系,异构同现模块用来学习模态间的相邻节点的关系。最后模型可以同时从模态内和模态间聚集邻域特征,学得更好的特征表示,从而提升模型的跨模态检索性能。Zheng等[56]设计了一种深度跨模态检索框架,该框架设计了一种新的分类损失(Instance loss)来挖掘模态内的细微差异,并将这种分类损失引入双路卷积神经网络中,构建了一种端到端的跨模态检索模型,该模型不仅能够学习图像和文本对中的细粒度跨模态信息,并且可以考虑模态内的数据分布,从而提取到了更丰富的图像和文本特征,并提高了模型图文检索的准确率。

受到生成对抗学习的启发,He等[51]提出了一种基于生成对抗学习的无监督跨模态检索方法,该方法使用2个前馈神经网络(生成器)作为图像和文本的特征提取器,将不同模态特征映射到公共特征空间中,然后使用模态分类器(鉴别器)来识别映射特征的模态信息。通过训练生成器和鉴别器,模型可以直接比较初始特征和映射特征的差别,确保不同模态的映射特征保留其原始模态特征信息,提高了模型的特征提取能力。和文献[54]类似,Chen等[57]提出了一种无监督生成对抗学习模型,该模型使用特征编码器(生成器)将单模态特征投影到公共特征空间中,模态分类器(鉴别器)用来区分不同模态的特征,生成器试图通过最大化其输出信息熵来欺骗鉴别器,鉴别器利用其输出概率计算香农信息熵来衡量其模态分类的不确定性。然后模型通过信息熵最大化来减少模态内特征的分布差异,减少特征提取时的信息损失,该模型同时利用KL(Kullback-Leibler)散度和双向三元组损失来减小共享空间中模态间特征的语义差异。

除了上述方法外,注意力机制[15]可以聚焦重要信息,并同时具备不同特征空间以及全局范围内的特征聚合能力,可以帮助模型提取到不同模态间更有关联的特征,因此注意力机制的跨模态检索方法是目前的主流方法。例如,Lee等[16]设计了一种基于交叉注意力的跨模态检索模型(Stacked cross at-

tention, SCAN),如图3所示。该方法使用交叉注意力将不同模态特征的相似性计算分为图像到文本和文本到图像,以图中展示的图像到文本相似性计算过程为例,首先将提取到的图像特征和文本特征映射到公共特征空间,然后将图像区域特征和文本特征使用交叉注意力来增强文本中不同单词特征之间的语义关联,最后累加所有图像区域特征与文本特征之间的相似性作为图像到文本的相似性。文本到图像的相似性计算则是计算所有单词特征和图像特征的相似性,最后模型通过最大化上述两种相似性来进行模型训练。为了充分使用注意力机制来对齐语义信息,Ji等[17]提出显著引导注意网络模型(Saliency-guided attention, SAN),该模型使用不对称的视觉注意力模块和文本注意力模块来学习视觉和语言之间的细粒度关联性。具体来说,SAN模型包括3个组件:显著性检测器、显著性加权视觉注意模块和显著性引导文本注意模块,其中显著性检测器利用图像数据提供视觉上的显著性信息来驱动2个注意力模块,显著性加权视觉注意模块能够学习更多区分性视觉特征,显著性引导文本注意模块将文本与视觉特征进行融合来保留与图像信息相关的文本特征,最后最大化文本特征和图像特征的相似性来训练模型。但是在该方法中,当文本中含有多义词汇时,模型在注意力机制的作用下会将多义词计算成均匀的特征值,这种值通常不具备语义信息,使得与其他样本特征之间的关系非常模糊。针对这个问题,Song等[18]提出一种基于多头注意力机制的跨模态检索框架,该框架由两个多义词实例嵌入网络模型组成。多义词实例嵌入网络利用多头自注意力机制来关注输入实例的不同特征(区域、单词),从而获得多个局部表示,然后通过残差学习将每个局部表示和全局表示结合起来计算同一个样本的多个不同表达,使得模型对多义词具有较强的鲁棒性。

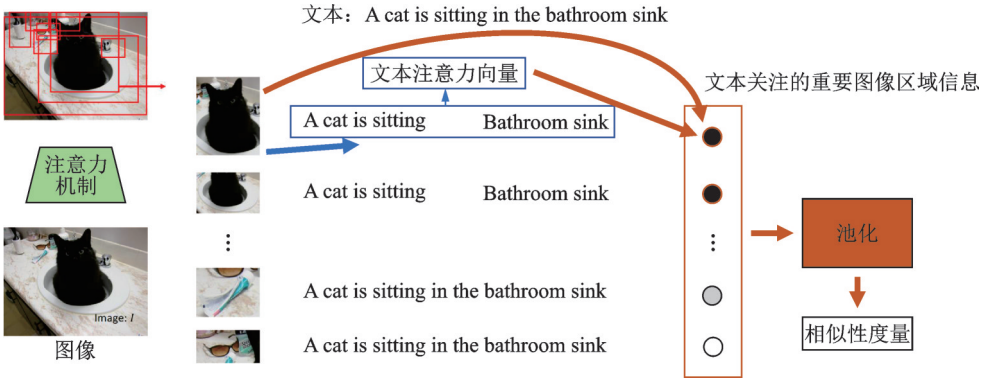


图3 SCAN模型结构图[16]

Fig.3 Structure diagram of SCAN model[16]

无监督深度学习方法由于没有使用手工标签来辅助训练,模型无法利用标签中的语义信息来对齐文本视觉特征高层语义。因此不少研究者通过将其他理论加入到模型中来解决这个问题,例如基于生成对抗学习的模型[54,57]可以通过鉴别器和生成器之间的相互对抗来保留模态内的信息分布,减少特征提取时的信息损失。基于注意力机制的模型[16-18]可以通过在特征提取与模态交互阶段引入注意力机制,来帮助模型实现对图文细粒度特征的提取和特征间的细粒度对齐。无监督深度学习模型不使用昂贵的手工标签信息,因此在未来仍然具有巨大的发展空间。

2.2.2 有监督学习方法

有监督方法通常利用标签信息在训练过程中监督模型从而获得更好的特征表示。Li等[58]提出一种基于深度学习的跨模态检索方法。该方法使用2个独立的卷积神经网络将同一样本的不同模态特征投影到公共特征空间中,该特征空间的维度与标签的类别数量一致,再将样本对应的标签转化成向量

表示。通过最大化样本不同模态投影特征与标签向量的相似度，让模型能够学到更细粒度的特征表示。为了有效地提取图像特征，Wei等[59]提出深度语义匹配方法，通过非对称构造的编码器来对文本和图像进行特征提取。具体来说，图像使用在大规模数据集上经过预训练的卷积神经网络作为图像的特征编码器，文本使用全连接神经网络作为编码器进行特征提取，将2个编码器的最后一层的维数设置为标签类别数，最后在公共特征空间中和对应的标签向量进行相似性度量，在标签信息的监督下，使得模型能够学习更具判别力的特征表示。Hu等[60]提出了一种可扩展深度跨模态检索模型。使用模态特定的编码器得到多模态数据的隐式表示，然后利用解码器对隐式表示进行解码。通过最小化原始特征和解码特征的距离减小样本不同模态特征编码时的特征信息损失，同时利用标签信息约束隐层表示来最大化样本不同模态特征之间的语义联系。由于每个模态都有独立的编码器，彼此之间互不影响，因此该方法可以在不重新训练整个模型的前提下，单独加入新的模态，使得模型具有很好的扩展性。现实情况下的数据往往带有噪声，会对跨模态检索方法的实际应用造成影响。为此Xu等[53]提出了一种应用于带噪声标签数据的跨模态检索对比学习方法，该方法通过对比学习将多模态数据投影到一个公共空间中，使用正则化处理来防止提取特征时噪声样本对模型的影响，并且使用动态权重平衡策略来缓解训练过程中聚类中心向噪声样本移动的问题，提高了模型对于噪声样本的抵抗力。Zhen等[52]提出深度监督跨模式检索方法(Deep supervised cross-modal retrieval, DSCMR)，如图4所示。与无监督方法[16]不同，该方法不仅在公共特征空间中对不同模态特征进行相似性比较，还引入线性分类器使得模型能够在标签空间中进一步对齐特征。具体来说，该方法使用特征编码器来提取图像文本特征，然后使用权值共享的全连接层将不同模态特征映射到公共特征空间中进行相似性的比较来消除模态间的差异。同时该方法将公共特征空间中的图像文本特征使用线性分类器生成对应的标签向量，并在标签空间中最小化标签向量和真实标签之间的距离，从而学习模态不变性特征。

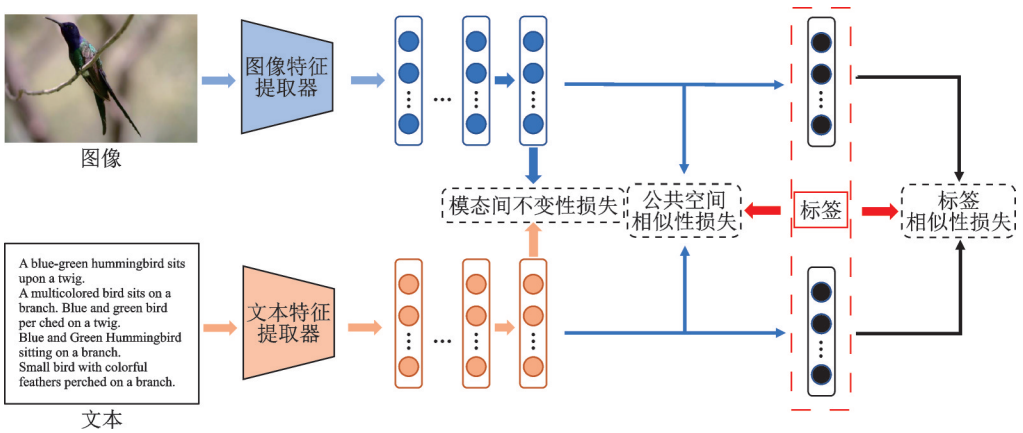


图4 DSCMR模型结构图[52]

Fig.4 Structure diagram of DSCMR model[52]

目前，监督深度学习方法能够利用标签信息使多模态样本学习具有判别力的特征。但是对于监督深度学习的科研工作仍然存在需要改进的地方：常用数据集里面包含的数据量和数据标签类别较少，如Wikipedia数据集[61]仅包含2 866个图像文本对，29种类别标签。在大数据的时代背景下，制作一个数据量大、样本类别充足的数据集是今后研究工作的重点方向。

2.2.3 自监督学习方法

自监督学习不需要标注成本高的人工标签，而是通过辅助任务根据数据本身的特性挖掘监督信息，已经广泛应用于跨模态检索领域。例如，Liu等[62]提出了自监督相关学习方法(Self-supervised cor-

relation learning, SCL),该方法将样本不同模态特征投影到公共特征空间中,将不同模态特征的语义信息作为监督信息,通过最大化原始图像特征和投影文本特征以及原始文本特征和投影图像特征之间的互信息,构建了原始特征和共同表征之间的内在关联,让模型学习到了更好的特征表示。随着Transformer在自监督学习领域的兴起,越来越多的研究者开始尝试将Transformer应用到跨模态检索任务中。基于Transformer跨模态检索的基本思路是:首先在大规模数据集上进行随机掩膜预测任务(Masked language model, MLM)训练,该任务首先随机掩盖一定比例的特征信息,将掩盖之前的完整特征作为监督信息。然后通过Transformer结构中全局注意力的作用,去预测被掩盖的信息,使得经过自监督训练出的特征能够更多地包含上下文信息,然后将模型在下游的检索任务中进行微调。为了在预训练过程中使模型感知到图像和文本的局部语义信息,使用图像特征提取器和文本特征提取器来提取图像的区域特征和文本的单词级特征,然后将图像和文本的局部特征拼接起来作为模型的输入进行自监督预训练。比如,Jiasen等[21]提出了基于Transformer的ViLBERT(Vision-and-language BERT)预训练模型,通过将BERT架构扩展为多模态双流模型,使用2个BERT模型分别处理文本和图像的特征,将2个BERT模型中的自注意力模块的键值对相互替换,可以将文本关注的视觉特征融入语言表现中,反之亦然。在真实场景下,图像中的物体往往存在重叠的情况,对于重叠区域的较大的不同物体,使用区域特征提取器提取的特征会非常相似,这将会对模型产生干扰。为了解决重叠情况下的语义相似问题,Li等[22]提出了Oscar(Object-semantics aligned)多模态预训练模型,如图5所示。该模型将基于预训练的图像特征提取模型提取到的图像局部信息的文本标签作为图像标签,将文本的单词信息、图像的标签信息和图像的区域信息连接起来作为预训练模型的输入,然后在文本信息中随机掩盖掉一部分信息来进行预训练。与基于有监督的方法[54]不同,该模型将被掩盖的单词特征和通过图像特征提取器获得的物体标签作为自监督信息,使得模型能够在自监督信息的帮助下区分具有较大重叠区域的物体特征,因此模型能够更好地与文本特征进行语义对齐和学习高层语义信息。也有研究者对Transformer模型的预训练任务进行改进,让模型更好地适应跨模态任务。比如Yu等[19]提出了一种基于知识增强的跨模态检索预训练框架(Knowledge enhanced vision-language, ERNIE-ViL)。该方法通过提取文本数据的名词、形容词以及关系词(如空间位置、动作等)构建场景图以获得文本的结构化信息,将其与图像提取到的所有局部特征连接起来输入到模型中,在预训练时分别进行对象掩膜预测、属性掩

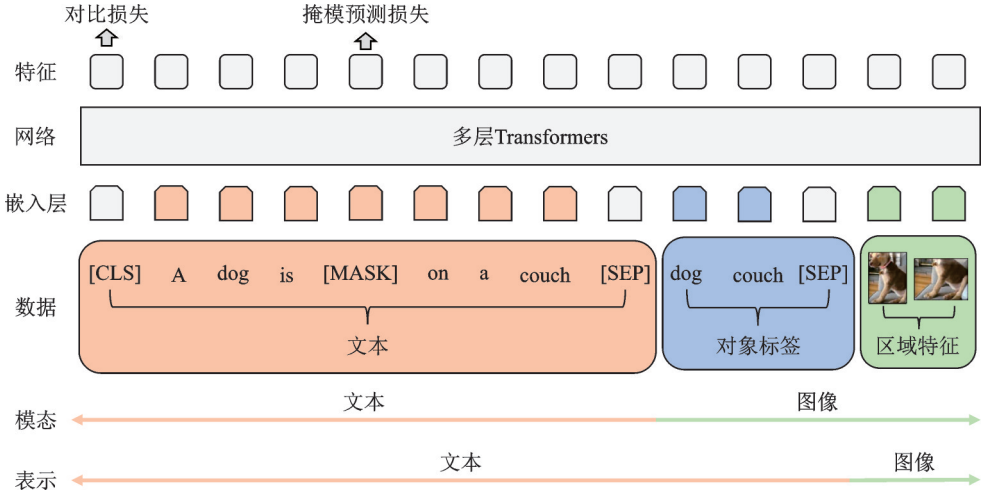


图5 Oscar模型结构图[22]

Fig.5 Structure diagram of Oscar model[22]

膜预测和关系掩膜预测任务。其中对象掩膜预测使得模型能够在对象级别进行视觉语义对齐;形容词掩膜预测提升模型对不同对象特征的识别能力;关系掩膜预测让模型区分相同对象之间不同关系的能力。Li等[23]提出了基于对比学习的预训练模型(Towards unified-modal, UNIMO),该模型利用大规模的单模态数据和多模态数据来进行对跨模态比学习从而提高其视觉和文本理解能力。Geigle等[63]认为目前基于Transformer的跨模态检索方法存在巨大的检索延迟和低效问题。为了解决这个问题,作者在Oscar[22]的基础上作出改进,使用权值共享的Transformer模块处理图像输入、文本输入以及联合图文的输入。在训练过程中将文本特征和图像特征连接起来输入Transformer模块中进行特征融合,将得到的特征输入到判断器中判断是否匹配。在检索过程中,首先使用训练好的Transformer模块提取所有图像和文本的特征,并且将特征向量保存到本地,然后通过两阶段的步骤实现检索任务。以图像检索文本为例:第1阶段计算该图像特征与测试集中所有待检索文本特征的相似度,得出前k个相似文本特征对应的文本;第2阶段将该图像分别与最相似的k个文本进行拼接,逐步输入到Transformer模块得到更准确的匹配分数。由于第1阶段使用的图像特征和文本特征只需要处理一次,图像和文本不需要反复使用参数庞大的Transformer进行特征提取,因此两阶段的检索方式极大地提升了检索效率。

上述基于自监督学习的跨模态检索方法虽然在检索精度上得到了很大的提升,并且文献[63]对检索效率进行了优化,但是仍然面临神经网络训练参数量大、训练的时间复杂高的问题,难以部署到实际应用中。因此,如何在确保模型稳定性的情况下降低时间复杂度有巨大的研究价值。

2.3 基于哈希表示的方法

多媒体数据量的急剧增长,研究人员开始采用跨模态哈希方法来解决跨模态检索问题,通过将多模态数据转化为二进制编码,投影到公共汉明空间,在提高检索速度的同时缩减存储空间。相较于实值编码,哈希编码方法提升了效率,从而被广泛应用于各种检索研究。目前提出的基于哈希的方法根据是否使用标签信息可以分为两类:(1)无监督学习方法,(2)有监督学习方法。表3中列出了具有代表性的方法,同时为了直观地展示不同类别方法之间的区别,从表3中分别选出基于无监督、有监督的2个模型进行结构展示。

表3 代表性基于哈希表示方法简要介绍

Table 3 Brief introduction to representative Hash based representation methods

时间	作者	代表方法	优势	局限性
2014年	Ding等	CMFH[25]	使用潜在因子模型集体矩阵分解来学习统一的哈希代码,能有效减小模态间的语义鸿沟。	该方法假设同一样本不同模态的哈希码完全一致,因此在进行语义对齐的过程中会丢失部分特征。
2016年	Tang等	SMFA[64]	使用集合矩阵分解获得潜在的语义特征,能够考虑不同模态之间的标签一致性和单个模态中的局部几何一致性。	模型对数据变化鲁棒性差,即对于新出现的模态数据不能生成高分辨率的哈希码。
2017年	Jiang等	DCMH[65]	通过监督信息来拉近不同模态哈希码的距离,能够保留各自模态的特征信息。	不能很好地处理数据的模态内相关性并且在训练过程中不同模态的哈希码之间失去了关联性。
2019年	Su等	DJSRH[66]	通过构造一个能够整合不同模态邻域信息的联合语义矩阵,使得模型能够获得不同模态特征中的高级语义信息。	没有考虑模态内部的潜在关系,在挖掘深层相关性时无法保留重要的语义信息。
2022年	Zhang等	TSDH[67]	将任务拆分为哈希码生成和训练两个部分,提升了哈希码的辨别性。	特征提取和模型训练被分离开来,使得模型效果容易受到特征提取模块的影响。

2.3.1 无监督学习方法

无监督哈希方法利用同一样本不同模态之间的对应关系学习模态一致的哈希编码,再利用样本内部不同模态之间的对应关系将不同模态的哈希码彼此靠近。Ding等[25]提出了集合矩阵分解哈希方法(Collective matrix factorization Hashing, CMFH)。该方法假设同一样本的不同模态特征都具有相同的哈希码,基于这个假设CMFH使用潜在因子模型集体矩阵分解来构造不同模态特征的唯一哈希码,之后基于哈希码来为每个模态训练对应的哈希生成函数。由于该方法假设同一样本不同模态的哈希码完全一致,而图像特征相比文本特征其语义信息要丰富得多,因此在进行语义对齐的过程中会丢失图像的部分特征。为了获得图像的显著特征结构,Zhou等[26]提出了潜在语义稀疏哈希模型(Latent semantic sparse Hashing, LSSH)。该模型利用稀疏哈希编码来捕获图像的显著特征,使用矩阵分解来提取文本中的潜在语义信息进而提升模型的语义对齐能力。Zhu等[68]提出语义辅助哈希方法(Unsupervised visual Hashing with semantic assistant, SAVH)。该方法具有2个子任务:第1个子任务通过构建主题超图来建立图像的高阶语义关系;第2个子任务通过矩阵分解来关联图像和检测到的潜在主题。通过联合两个子任务使得模型可以从噪声文本中提取不受影响的语义信息来增强哈希码的鉴别能力。Su等[66]将深度神经网络和哈希方法结合,提出了深度连接语义重构哈希方法(Deep joint-semantics reconstructing Hashing, DJSRH),如图6所示,该方法通过拉近联合语义相似矩阵和跨模态编码矩阵之间的距离来提升模型的语义对齐能力,与基于深度学习的方法[16]不同,该方法将不同模态的特征映射到哈希空间而不是实值空间。具体来说,该方法首先在模态内部聚合原始邻域信息来构造联合语义矩阵,之后将不同模态特征通过哈希生成器映射到公共哈希空间中,并在公共哈希空间中构造跨模态编码相似度矩阵,最后在公共空间中最小化这2个矩阵之间的距离,使得模型能够获得不同模态特征中的潜在高级语义信息,并且能够在学习哈希码的同时保持原始数据的邻域结构不变性。Zhang等[69]提出无监督多路径对抗生成哈希方法(Muti-pathway generative adversarial Hashing, MGAH)。该方法基于生成对抗原理,生成器通过给定样本的特定模态特征来拟合与给定信息分布相同的虚假特征并联合该样本另一模态对应的真实特征数据来欺骗鉴别器,鉴别器通过区分生成的假数据和从其他模态中采样的真数据来互相对抗以促进哈希函数的学习。但是此方法没有考虑模态间的潜在关系,在挖掘深层相关性时无法保留重要的语义信息。为此,Zhang等[70]提出了注意引导语义哈希方法(Aggregation-based graph convolutional Hashing, AGSH),该方法使用注意力机制构建注意感知语义融合矩阵,该矩阵集成了来自不同模态的重要特征,进而掌握了不同模态的重要语义信息,通过最小化哈希码和融合矩阵之

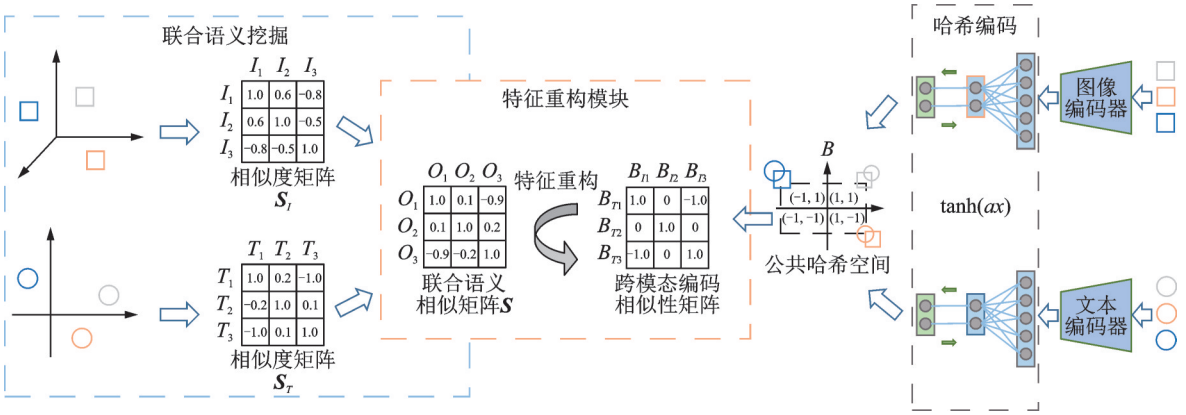


图6 DJSRH模型结构图[66]

Fig.6 DJSRH model structure diagram[66]

间的误差来训练模型。模型在实际应用中会出现数据丢失的情况，这将导致多模态部分数据缺少成对信息，从而影响模型性能，为此 Chen 等[71]提出了无监督深度哈希方法（Unsupervised deep imputed Hashing, UDIH）。该方法训练过程由 2 个阶段组成：首先将不成对的数据由其提出的生成器进行插补；然后在相关图上应用一个具有加权三重损失的神经网络来学习每个模态在二进制空间中的哈希码，能够使模型保持不同模态数据之间的语义一致性和差异性。

无监督哈希方法仅使用样本内不同模态数据的对应性来进行训练，因此适用于大规模且数据分布均匀的数据集。此类方法在处理跨模态异构问题上表现出了性能优势，但是因为缺乏监督信息，在学习过程中会造成信息损失，因此减少信息损失是目前需要研究的重要内容。

2.3.2 有监督学习方法

有监督哈希方法能够利用标签信息辅助模型的学习，从而得到更具代表性的哈希码。例如 Jiang 等[65]提出了一种深度跨模态哈希框架（Deep cross-modal Hashing, DCMH），如图 7 所示。该框架将特征学习和哈希生成集成到一起，构建了一个端到端的深度哈希图文检索模型。该模型通过哈希生成模块来生成各自模态的哈希码，与基于无监督的哈希方法[25]不同的是，该方法没有在公共哈希空间中比较不同模态哈希码之间的相似性，而是根据监督信息来最大化两种模态特征的相似性，并且通过距离最小损失拉近模态特征及其哈希码的距离，能够与让哈希码最大化保留各自模态的特征信息。但是 DCMH 不能很好地处理数据的模态内相关性，并且在训练过程中不同模态的哈希码之间失去了关联性。为此 Meng 等[72]设计了一种具有离散表示的非对称学习方法，该方法将不同模态的映射矩阵分解为公共表示矩阵和模态特定矩阵，其中公共表示矩阵挖掘了模态间共享的语义表示，模态特定矩阵捕获了每个模态的特定属性，通过这种方式模型能够更准确地提取同一样本的不同模态表示之间共享的内在语义。该方法使用语义标签来指导哈希码学习过程，充分利用了监督信息，使学习到的哈希码的辨别能力得到了很大的提高。Liang 等[73]提出了双向注意力生成对抗哈希方法（Dual-pathway attention based supervised adversarial Hashing, DASAH），该方法使用双向注意力机制来让模型获得细粒度的模态特征，将特征提取和生成对抗哈希学习集成在一个统一的框架中进行联合学习和优化，进一步提高了模型的性能。Zhang 等[67]提出了两阶监督离散哈希（Two-stage supervised discrete Hashing, TSDH）方法。该方法在第 1 阶段将样本的不同模态特征通过哈希构造函数生成哈希码；第 2 阶段将哈希码直接与标签向量进行距离最小化训练，通过这两阶段的学习使得模型能够生成更有辨别性的哈希码。为了有效利用标签信息和特征的空间位置信息，Tang 等[64]提出了基于矩阵分解的跨模态哈希方法（Su-

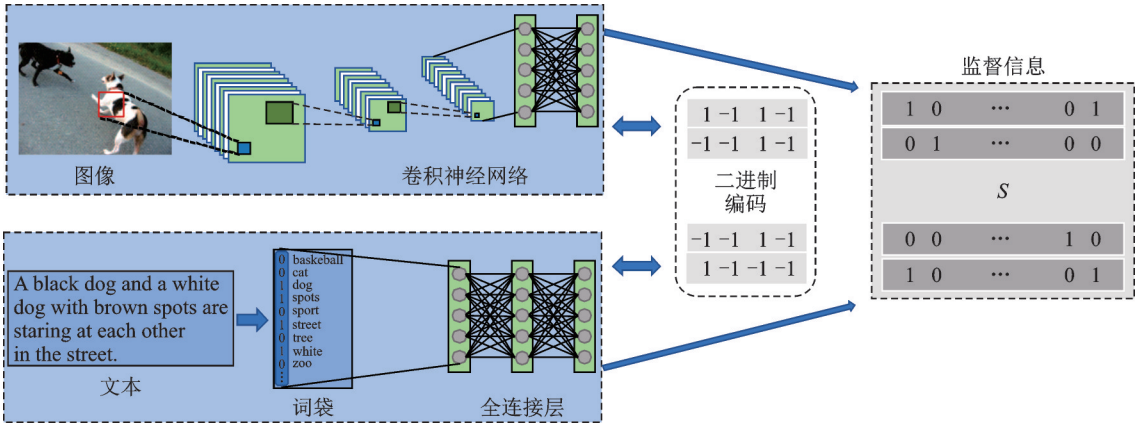


图 7 DCMH 模型结构图[65]

Fig.7 DCMH model structure diagram[65]

pervised matrix factorization Hashing, SMFA)。该方法使用矩阵分解来获得不同模态数据内部的高级语义特征,并且在训练过程中保持了不同模态之间的标签一致性和单个模态中的局部几何一致性。但是该方法不能很好地适应数据的变化,即对于新出现的多模态数据,不能生成高分辨率的哈希码。为了解决这个问题,Chen等[74]提出了基于有监督矩阵分解和图正则哈希方法(Intra-and inter-modality similarity preserving Hashing, IISPH),该方法使用更灵活的哈希函数,在生成哈希码的同时保持相同样本不同模态之间的相似性和低维特征空间中每个模态的局部几何结构,然后利用监督标签信息在低维公共空间中细化局部邻域结构,使得哈希函数在学习过程中能够保持原始特征空间的全局和局部信息。Li等[75]提出了有监督鲁棒离散多模态哈希(Supervised robust discrete multimodal Hashing, SRDMH)。该模型直接将标签信息引入哈希函数的学习中,使得模型能够生成更有辨别力的哈希码,进而提升模型的精度。

基于监督的哈希方法由于利用了标签信息通常能够获得更好的检索精度,并且标签信息在数据特征提取和哈希函数学习中可以灵活使用,缓解了使用二进制编码表示特征导致模型精度不高的问题,然而在监督条件下模型需要使用昂贵的手工标签信息,因此在未来的研究中可以考虑使用辅助任务根据数据本身的特性挖掘监督信息,从而以自监督的方式来进行学习。

3 常用数据集和评价指标

在对相关综述总结的基础上,为了进一步加深对跨模态检索的认识与理解,评估与分析不同跨模态检索技术方法的特点,文本收集了目前跨模态检索性能的评价指标以及该领域常用的公开数据集。下面将详细介绍公开数据集以及评价指标,并对部分数据集上现有方法的性能表现进行详细分析。

3.1 数据集介绍

跨模态检索的常用数据集如表4所示,其中类别表示数据集中具有多少种类别标签,各数据集具体介绍如下:

表4 常用跨模态数据集比较

Table 4 Comparison of commonly used cross modal datasets

数据集	类别	图像数量/个	文本数量/个
MS-COCO[27]	91	173 791	868 955
Flickr-30K[76]	—	31 783	158 915
Wikipedia[61]	29	2 866	2 866
NUS-WIDE[77]	81	269 648	269 648
INRIA-Websearch[78]	353	71 478	71 478
IAPRTC-12[79]	—	19 627	74 961

(1)MS-COCO[27]。该数据集是微软构建的大型图文数据集,是跨模态检索最常用的数据集之一。该数据集中的图像内容来自与人们的日常生活,数据集中的图像总共包含91个类别(检测任务使用80类),共包含有123 287张图像,其中每张图像对应5个文本描述。

(2)Flickr-30K[76]。该数据集是由雅虎构建的图文数据集,该数据集经常和MS-COCO数据集一起用于图文跨模态检索中。其中的图像文本来源于Flickr网站,共包含31 783张图像,与MS-COCO类似,每张图像具有5个对应的文本描述,每一张都有5个文本注释,因此总共具有158 915个图像文本对。其中文本和图片主要是描述参与各种日常获得的人物信息。

(3)Wikipedia[61]。该数据集由维基百科中的特色文章组成,该文章由2 866个图像文本对组成,每张图像对应一个文本描述,其中文本是一篇描述人物、地点或某些事件的文章,而图像则是文章中的插图,该数据集总共有29个概念类,其中10个为主要概念类。该数据集与其他数据集相比不仅数据量小而且类别数量有限。

(4)NUS-WIDE[77]。该数据集是由新加坡国立大学构建的图文数据集,该数据集经常和Wikipedia数据集一起使用,其图像主要来源于Flickr网站,包含有269 648张图像,每张图像对应一个文本描述,

因此总共具有269 648个图像文本对。

(5)INRIA-Websearch[78]。该数据集由法国的Lear研究团队构建,由71 478个图像文本集合组成,一共有353个类别。注释文件由与图像相关的手动标签和其他相关元数据组成,例如网页URL、图像URL、页面标题等信息,数据集中的图像被缩放到150像素×150像素的正方形,该数据集中仅包含图像和文本信息。因此常用与图文跨模态检索任务中。

(6)IAPRTC-12[79]。该数据集最初由Grubinger等构建,共包含有19 627张图像,每个图像对应有1~5个文本描述,该数据集中的文本具有的总词汇量为4 424。该数据中的文本描述都符合语法要求并且图像描述详尽,因此数据集中几乎没有噪音。

3.2 评价指标

跨模态检索包含两种子任务:(1)用图像作为查询去检索与之相关的文本(Image to text);(2)用文本作为查询去检索与之相关的图像(Text to image)。在测试阶段,给定一个文本或图像查询样本,检索出语义相关的图像或文本。评估时考虑的2个典型因素是:(1)查询和结果之间的类相关性评估;(2)检查图像文本对的跨模态相关性。第1个因素说明模型学习不同模态潜在表征的能力,而第2个因素说明模型学习不同模态相关潜在概念的能力。与上述两个因素相关的指标如下:

3.2.1 平均精确率均值

平均精确率均值(mean Average precision, mAP)常用于Wikipedia和NUS-WIDE等数据集中,该指标衡量检索到的结果是否与查询数据属于同一类(相关)或不属于(无关)。它是所有查询计算的平均精度的平均值。给定查询样本q以及关于该样本的R个检索结果,平均精确率(Average precision, AP)可以被定义为

$$\mathrm{AP}(q)=\frac{1}{R}\sum_{r=1}^{R}P(r)\mathrm{rel}(r) \tag{5}$$

式中:r表示检索结果中的第r个样本;$P(r)$表示第r个检索样本的精确度;当r与查询样本相关时,$\mathrm{rel}(r)$结果为1,否则$\mathrm{rel}(r)$为0。此后,基于平均精确率可以计算得到最终mAP的值,公式为

$$\mathrm{mAP}=\frac{1}{Q}\sum_{q=1}^{Q}\mathrm{AP}(q) \tag{6}$$

式中:Q代表查询样本数;q表示其中的一个样本。mAP值越大,表示跨模态检索的准确率越高。

3.2.2 召回率

召回率(Recall)表示为检索模型检索出的与查询样本相关样本与数据集中与查询样本相关的比值,定义为

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{7}$$

式中:TP表示与查询结果相关的结果数量;FN表示没有没检索到的与查询样本相关的样本数量;TP+FN表示数据集中与查询样本相关的样本总数量

在跨模态检索中与召回率相关的常用评价指标还有Recall@K, $K\in[1,5,10]$,表示为R@1、R@5和R@10,分别表示在前1、5和10个结果中检索到的真值的百分比,值越高表示模型的性能越好,定义REL_k代表前k项结果中的相关项数,REL表示给定查询的相关项总数,其计算公式如下

$$\mathrm{Recall@}K=\frac{\mathrm{REL}_k}{\min(k,\mathrm{REL})} \tag{8}$$

3.2.3 精确率

精确率(Precision)定义为检索模型检索出的与查询样本相关的样本与被检索到的所有样本的比

值,定义如下

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{9}$$

式中:FP表示与查询样本不相关样本的数量;TP+FP表示检索到的总结果数量。

3.2.4 精确率-召回率曲线

精确率-召回率曲线(Precision-recall curve)代表了模型精确率和召回率之间的关系,将模型的召回率设置为横坐标,精确率设置为纵坐标,该指标将模型对应的曲线下的面积大小作为该模型的性能,如果一个模型的曲线被另外一个模型产生的曲线所包含,则可以认为后者的性能要优于前者。

3.3 实验分析

根据表4给出的信息,可以将这些数据集分为2类:不对称数据集和对称数据集。MS-COCO、Flickr-30K和IAPRTC-12都是一个图像对应多个文本,因此将它们归为不对称数据集;而Wikipedia、NUS-WIDE和INRIA-Websearch都是一张图像对应一个文本,所以将它们归为对称数据集。下面将从对称数据集和不对称数据集中分别选择2个数据集进行实验分析,分别是不对称数据集中的MS-COCO和Flickr-30K以及对称数据集中的Wikipedia和NUS-WIDE。

3.3.1 Wikipedia和NUS-WIDE数据集上相关方法分析与比较

本节在Wikipedia和NUS-WIDE数据集上对传统方法、基于深度学习的跨模态检索方法和基于哈希表示的跨模态检索方法进行性能上的分析与比较。

表5,6给出了较为先进的基于传统方法和基于深度学习的图文跨模态检索方法的相关结果,表中Avg表示前两项的均值。由于大部分图文跨模态检索方法都会结合Wikipedia和NUS-WIDE一起使用,因此将表5和表6放在一起进行比较。其中,CCA方法[32]利用矩阵分解将不同模态数据映射到公共特征空间中,然后在这个空间中最大化它们的相似性;UCAL方法[55]基于对抗学习思想,通过对抗训练在保留模态特征信息的同时学习模态间的公共表示;SCL方法[62]将提取的特征中的语义信息作为标签信息。通过最大化原始图像特征和投影文本特征、原始文本特征和投影图像特征之间的互信息,构建了原始特征和共同表征之间的内在关联,让模型学习到了更好的特征表示;DCCA方法[10]将CCA与深

表5 不同方法在Wikipedia数据集上的mAP比较

Table 5 Comparison of different methods in mAP on Wikipedia dataset %

方法名称	mAP		
	Text to image	Image to text	Avg
CCA[32]	13.4	13.3	13.4
UCAL[55]	26.3	27.3	26.8
SCL[62]	43.1	38.6	40.9
DCCA[10]	44.4	39.6	42.0
JRL[80]	44.9	41.8	43.4
ACMR[81]	47.9	42.6	45.2
CMDN[82]	48.8	42.7	45.8
CCL[83]	50.4	45.7	48.1
P3S[36]	52.0	46.9	49.5
DSCMR[52]	52.1	47.8	49.9
AAEGAN[84]	59.5	54.6	57.0

表6 不同方法在NUS-WIDE数据集上的mAP比较

Table 6 Comparison of different methods in mAP on NUS-WIDE dataset %

方法名称	mAP		
	Image to text	Text to image	Avg
CCA[32]	37.8	39.4	38.6
SCL[62]	50.8	46.9	48.9
CMDN[82]	49.2	51.5	50.4
CCL[83]	50.6	53.5	52.1
DCCA[10]	53.2	54.9	54.0
ACMR[81]	54.4	53.8	54.1
P3S[36]	54.5	57.4	56.0
JRL[80]	58.9	59.8	59.2
DSCMR[51]	61.1	61.5	61.3
ALGCN[85]	74.7	75.8	75.3

度学习结合起来,使用卷积神经网络处理非线性问题;JRL(Joint representation with sparse and semi-supervised regularization)方法[80]学习不同模态的稀疏投影矩阵来提升对噪声的鲁棒性;ACMR(Adversarial cross-modal retrieval)方法[81]利用标签信息最小化具有相同语义标签不同模态特征之间的距离,以及最大化不同语义标签不同模态特征的距离,提高了模型的性能;CMDN(Cross-media shared representation by hierarchical learning with multiple deep networks)方法[82]将模态特征信息映射到公共特征空间,然后将公共特征空间中的模态特征进行分层组合来进一步学习模态间语义相关性;CCL(Cross-modal correlation learning)方法[83]采用多任务学习策略来自适应的平衡模态内和模态间相关性,从而得到更精确的公共表示;P3S方法[36]利用共享子空间和私有子空间之间的正交性约束,使得方法能排除图像中不相关的场景信息或者文本中错误语句产生的干扰,从而让模型获得更具代表性的特征;DSCMR方法[52]在特征空间中最小化监督信息和对应不同模态特征之间的辨别损失,从而学习模态不变性特征,提升模型的性能。AAEGAN(Assembling auto-encoder and generative adversarial network)方法[84]通过相互重建样本模态数据,以类嵌入作为重建过程中的辅助信息,使跨模态分布差异最小化;ALGCN(Adaptive label-aware graph convolutional networks)方法[85]能够保留样本不同模态之间的跨模态语义关联,挖掘标签的深层结构信息,进而提升模型的精度。

从表5,6中可以看出相同的方法在这两个数据集上的准确率相差较大。相较于Wikipedia,NUS-WIDE有更大的数据规模和更丰富的类别信息,所以模型能够在NUS-WIDE数据集上进行更为充足的训练。下面以NUS-WIDE数据集为例,对涉及方法进行具体分析:相比于传统的CCA,基于深度学习的跨模态检索方法在性能上有很大提高,说明深度神经网络非线性投影特性对于跨模态检索有重要的作用;ACMR方法通过标签信息学习跨模态的一致性,相较于无监督的方法(如CMDN、CCL、DCCA)性能有了进一步的提升,这是因为ACMR正确利用了标签信息,从而得到了更具有判别力的多模态特征;相比上述方法,P3S性能具有明显提升,说明利用共享子空间和私有子空间之间的正交性约束能够有效地去除检索无关的噪声信息;DSCMR通过共享策略学习模态不变特性,保持相同样本下不同模态的语义相关性,相较于P3S,性能有了进一步提升,说明设计合理的学习策略,使模型学习到跨模态样本的不变语义有利于优化相关性的度量;ALGCN方法则通过保留模态关联来减小样本不同模态特征信息在投影到公共特征空间时损失的局部信息,同时使用样本标签的深层结构信息来拟合模态差距,因此在表6中得到了最好的效果。

表7,8给出了较为先进的基于哈希表示的图文跨模态检索方法的相关结果,其中16 bit、32 bit和64 bit表示哈希码的长度。表中:HFCVSS(Hash functions for cross-view similarity search)方法[86]通过最小化不同模态的平均汉明距离来学习哈希函数,但是该方法没有考虑到模态之间的差异,因此检索性能有限;IMH(Inter-media Hashing)方法[87]思想与HFCVSS方法相似,但是该方法充分考虑了模态之间的关联和差异,并且保持彼此最近的样本的模态间和模态内相似性,但是IMH需要计算大量样本的相似度图因此检索效率低,不适合大规模数据集;CMSSH(Cross-modality metric learning using similarity-sensitive Hashing)方法[88]作为最初的有监督哈希方法,通过根据样本的相似性来生成正负样本对,然后构造两组线性

表7 哈希表示方法在Wikipedias数据集上的mAP比较

Table 7 Comparison of hash representation methods in mAP on Wikipedia dataset %

方法名称	Image to text			Text to image		
	16 bit	32 bit	64 bit	16 bit	32 bit	64 bit
HFCVSS[86]	12.57	12.12	12.15	11.85	10.34	10.24
IMH[87]	15.73	15.75	15.68	14.63	13.11	12.90
CMSSH[88]	18.77	17.71	16.46	16.30	16.17	15.39
CMFH[25]	21.32	22.59	23.62	44.84	51.32	52.69
LSSH[26]	21.41	22.16	22.18	50.31	52.24	52.93
HSCM[89]	22.10	23.37	24.42	21.35	23.66	24.79
SePH[90]	27.83	29.56	30.49	63.18	65.77	66.46
TSDH[67]	36.41	38.56	39.08	75.68	77.03	77.31

哈希函数作为分类器,将检索损失转化为了分类损失;CMFH方法[25]假设同一样本不同模态数据的哈希码在映射到汉明空间中是一致的,将样本不同模态的数据使用矩阵分解来学习统一的哈希代码,之后基于哈希码来为每个模态训练对应的哈希生成函数,进而提高了该模型的检索性能;LSSH方法[26]是CMFH方法的扩展,该方法在CMFH的基础上使用稀疏编码来获得更图像高级特征信息,使用矩阵分解来学习潜在的语义信息,然后将生成的信息映射到公共空间中进行相似性的计算,因此LSSH方法相比于CMFH方法在性能上具有一定的提升;HSCM(Hashing with semantic correlation maximization)方法[89]是一种有监督的哈希方法,该方法利用标签信息学习公共空间的表示,具有较低的时间复杂度;SePH(Semantics-preserving Hashing)方法[90]使用从模态数据中提取的语义相关性矩阵作为监督信息来学习哈希码,该方法模型复杂,需要较长的训练时间;STMH(Semantic topic multimodal Hashing)方法[91]通过对文本数据进行聚类来获得隐藏的文本主题,通过矩阵分解来挖掘图像数据的语义信息,并引入范数来增强矩阵分解的鲁棒性,使模型生成的哈希码具有较高的离散性,从而能够保存更多的信息;DCMH方法[65]使用哈希生成模块来生成不同模态样本的哈希码,根据监督信息来最大化两种模态间特征的相似性,再使用距离最小损失来拉近模态特征和其哈希码的距离,从而让哈希码最大化保留各自模态的特征信息;PRDH(Pairwise relationship guided deep Hashing)方法[92]与DCMH非常相似,该方法集成不同类型的成对约束来衡量模态间和模态内数据哈希码的相似性,并且该方法使用去相关约束来加强哈希码的离散性,因此其模型性能较DCMH有一定的提高;ADAH(Attention-aware deep adversarial Hashing)方法[93]引入注意力机制来加强模型的学习,模型使用注意力机制进行细粒度的语义对齐,但是对模型过度使用注意力机制可能会使模型忽略太多有效信息;TSDH方法[67]由两阶段组成:第一阶段将每个模态样本使用离散优化方法直接生成哈希码,减少了模型的量化误差;第二阶段将哈希码直接与标签向量进行距离最小化训练,增强了哈希码的区分能力。

表8 哈希表示方法在NUS-WIDE数据集上的mAP比较

Table 8 Comparison of hash representation methods in mAP on NUS-WIDE dataset %

方法名称	Image to text			Text to image		
	16 bit	32 bit	64 bit	16 bit	32 bit	64 bit
STMH[91]	47.10	48.64	49.42	44.71	46.77	47.80
CMFH[25]	49.00	50.31	50.97	50.31	51.87	52.25
LSSH[26]	49.33	50.06	50.69	62.50	65.78	68.23
HSCM[89]	54.09	54.85	55.53	53.44	54.12	54.84
DCMH[65]	59.03	60.31	60.93	63.89	65.11	65.71
SePH[90]	60.37	61.36	62.11	59.83	60.25	61.09
PRDH[92]	61.07	63.02	62.76	65.27	69.16	67.20
ADAH[93]	64.03	62.94	65.20	67.89	69.75	70.39
TSDH[67]	64.73	65.74	66.89	75.52	73.38	78.69

由表7,8可以看出,IMH通过考虑模态之间的关联和差异,因此得到了比HFCVSS更好的效果。LSSH方法是对CMFH的进一步改进,该方法使用稀疏编码来获取图像的高级特征信息,因此其效果在上述2个数据集中都高于后者。HSCM方法通过使用标签信息来监督训练过程,因此其效果相较基于无监督方法的LSSH以及CMFH具有一定的提升,但是由于HSCM存在较大的量化损失,因此其性能没有同样属于监督学习的SePH方法好。DCMH方法通过拉近特征和哈希码的距离来减少特征转换为哈希码时产生的量化损失,因此其性能相较于HSCM具有明显的提升。PRDH方法由于在DCMH引入了额外的去相关约束,增强了哈希码的信息量因此其模型性能要比DCMH模型的性能具有一定的提高。TSDH通过直接生成哈希码来减小量化误差,并且同时利用监督信息来监督模型的训练,提高了模型的性能,因此在表中都有最好的效果。同时从表7,8中还可以看出量化误差对于模型性能的影响,因此如何减小特征在转化为哈希码过程中的量化损失是需要思考的问题。

3.3.2 MS-COCO和Flickr-30K数据集上相关方法分析与比较

由于传统方法和基于哈希的跨模态检索方法MS-COCO和Flickr-30K上应用较少,因此本节将重

点放在基于深度学习的方法上进行分析与比较。不同方法在MS-COCO和Flickr-30K数据集上Recall@K比较如表9,10所示。表中:Dual-Path方法[56]构建了一种端到端的深度图文检索模型,该模型不仅能够学习图像和文本对中的细粒度跨模态信息,而且可以考虑模态内的数据分布;SCO(Semantic concepts and order)方法[94]将模态全局特征和局部特征组合在一起,用来平衡不同模态间的语义概念和上下文的相对重要性;SCAN方法[16]将提取到的图像和文本特征使用交叉注意力进行细粒度语义对齐,选取出相互有语义关联的特征,最后通过最大化图像特征和文本特征的相似度来进行训练,提升了模型的性能;CAAN方法[7]使用注意力来捕获潜在的模态内关联,从而保证了特征的分布不变性;VSRN(Visual semantic reasoning)方法[95]是一种基于注意力机制的跨模态检索模型,该模型使用注意力机制来捕获不同模态的显著性特征,然后使用图网络来构建特征之间的关系,帮助模型学习到特征中的空间信息;MMCA(Multi-modality cross attention network)方法[96]提出一种交叉注意力机制网络,该网络不仅学习单模态内部元素的关联,而且挖掘不同模态中元素之间的关联,然后将学习到的两种不同的关联统一到同一空间下用于匹配度分析;GSMN(Graph structured network for image-text matching)方法[97]提出了一种图结构化匹配网络,该网络将对象、关系和属性明确建模为短语,并通过对这些本地化短语进行匹配来共同推断细粒度的对应关系;SGRAF(Similarity graph reasoning and attention filtration)方法[98]使用基于向量的相似性表示来更有效地建模跨模态关联,并且使用注意力过滤模块来减少无意义信息的干扰;IMRAM(Iterative matching with recurrent attention memory)方法[99]对SCAN作出改进,在交叉注意力的基础上增加了记忆过滤模块来去除不不重要的信息,该模型将多个这样的交叉注意力模块叠加来进行不同模态间的细粒度对齐;GPO(Generalized pooling operator)方法[100]提出了一种自适应池化策略,为每一种池化操作学习重要性系数,从而挑选出最适应图文匹配的池化操作;Oscar模型[22]将图像中物体的标签信息加入到模型预训练任务中,在物体标签的监督下帮助模型区分具有较大重叠区域的物体特征,使得模型能够更好地与文本特征进行语义对齐;UNITER(Universal

表9 MS-COCO数据集上Recall@K($K\in[1,5,10]$)比较

Table 9 Recall@K($K\in[1,5,10]$) comparison on MS-COCO dataset %

方法	Image-to-text			Text-to-image		
	Recall @1	Recall @5	Recall @10	Recall @1	Recall @5	Recall @10
Dual-Path[56]	44.2	70.2	79.7	30.7	59.2	70.8
SCO[94]	69.9	92.9	97.5	56.7	87.5	94.8
CAAN[7]	75.5	95.4	98.5	61.3	89.7	95.2
VSRN[95]	76.2	95.6	98.5	61.7	89.1	95.0
MMCA[96]	74.8	95.6	97.7	61.6	89.8	95.2
GSMN[97]	78.4	96.4	98.6	63.3	90.1	95.7
SGRAF[98]	79.6	96.2	98.5	63.2	90.7	96.1
IMRAM[99]	76.7	95.6	98.5	61.7	89.1	95.2
GPO[100]	79.7	96.4	98.9	64.8	91.4	96.3
Oscar[22]	88.4	99.1	99.8	75.7	95.2	98.3

表10 Flickr-30K数据集上Recall@K($K\in[1,5,10]$)比较

Table 10 Recall@K($K\in[1,5,10]$) comparison on Flickr-30K dataset %

方法	Image-to-text			Text-to-image		
	Recall @1	Recall @5	Recall @10	Recall @1	Recall @5	Recall @10
Dual-Path[56]	52.2	80.4	88.7	37.2	69.5	80.6
SCO[94]	55.5	82.0	89.3	41.1	70.5	80.1
SCAN[16]	67.4	90.3	95.8	48.6	77.7	85.2
CAAN[7]	70.1	91.6	97.2	52.8	79.0	87.9
VSRN[95]	71.3	90.6	96.0	54.7	81.8	87.2
IMRAM[99]	74.1	93.0	96.6	53.9	79.4	87.2
SGRAF[98]	77.8	94.1	97.4	58.5	83.0	88.8
GPO[100]	81.7	95.4	97.6	61.4	85.9	91.5
UNITER[101]	85.9	97.1	98.8	72.5	92.4	96.1
ERNIE-ViL[19]	86.7	97.8	99.0	74.4	92.7	95.9
UNIMO[23]	89.7	98.4	99.1	74.7	93.4	96.1

image-text representation)模型[101]通过预训练能够更稳定地应用于具有联合多模态嵌入的异构下游任务中;ERNIE-ViL模型[19]通过构建文本场景图来进行预训练,帮助模型学习到了细粒度的语义关系;UNIMO模型[23]利用大规模的单模态数据来进行对比学习,能够学习到精确对齐的多模语义表示。

从表9,10可以看出,由于非对称数据集中图像和文本的数量差异,导致这2个数据集中的方法用文本搜索图像任务准确率和用图像搜索文本任务准确率差异很大,其中SCAN在模型中加入注意力机制,相比于不使用注意力机制的SCO方法极大地提升了模型精度,证明了注意力机制在跨模态检索中的优势。CAAN在SCAN的基础上使用注意力来捕获不同模态的显著特征,因此模型性能相较于前者具有一定的提升。VSRN在使用注意力的同时通过图网络来获取特征之间的位置信息,因此得到了比CAAN更好的效果。GPO使用自适应池化策略来为模型选取更有代表性的特征,其性能相比于使用注意力机制的IMRAM和SGRAF具有明显的提升,这证明了合理利用池化方法的重要性。UNIMO、ERINE-VIL和Oscar等预训练模型的性能在表9,10中都远高于其他方法,这足以说明预训练的方法具有挖掘潜力。跨模态检索方法分析与总结如表11所示。

表11 跨模态检索方法分析与总结

Table 11 Analysis and summary of cross modal retrieval methods

类别	实现思路	优势	劣势	总结	使用场景
传统方法	通过统计分析方法将不同模态的手工特征映射到公共特征空间,通过构造映射矩阵、矩阵分解等方法在特征空间中进行相似性度量。	使用统计分析方法来最大化模态间相关性,模型构造简单并且扩展性强,能够直接对不同模态数据进行建模,具有较强的可解释性。	传统方法对非线性关系处理能力差,虽然有不少方法通过核函数、高维映射等方式来处理非线性关系,但是会导致模型效率下降。	传统方法主要依靠统计分析来进行跨模态检索,模型结构简单,但是模型对模态内数据细粒度信息提取和模态间语义对齐专注较少,并且由于模态间的关系往往是非线性的,传统方法对表现不佳。	构造简单的小规模数据集
基于深度学习的方法	使用特征编码器来提取不同模态的特征,然后使用神经网络模型将不同模态数据映射到公共空间中进行相似性度量。	深度神经网络能够很好地处理非线性关系,并且在特征提取方法的表现要优于手工特征,借助深度神经网络的学习能力可以让模型学到更具代表性的多模态特征表示。	过于关注底层特征和深层网络相关性,忽视了模态内部的区域特征和不同模态之间的语义结构信息,模型结构复杂,可解释性差。	基于深度学习的方法在特征提取和语义对齐的方法具有很大的优势,但是模型的参数数量较大,在大规模数据集上需要耗费巨大的算力和时间。	常见的大规模数据集
基于哈希表示的方法	将不同模态的数据通过哈希生成函数映射到公共汉明空间中,并在此空间中进行相似性的度量。	基于哈希表示的方法将多模态数据转化为二进制编码,在提高检索速度的同时减小了储存空间,相较于基于深度学习的方法能够获得更高的检索效率。	在将模态特征转换为二进制编码的过程中会导致原有模态特征结构被破坏,导致部分信息丢失,因此模型检索精度相较于基于深度学习的方法表现较差。	基于哈希表示的方法具有检索效率高、储存空间小的优点,但是在二进制转换的过程中会丢失部分信息,同时没有考虑不同模态数据的结构差异。	常见的大规模数据集

4 研究展望

(1) 更好的可塑性和稳定性

人类有能力将一个任务学到的知识运用到另一项任务中,学习新任务后也不会忘记如何完成前一个任务,即人类有很强的可塑性(学习新知识的能力)和稳定性(保留旧知识的能力)。但是,对于跨模态检索模型,由于其自身的设计天然存在灾难性遗忘问题。比如,使用食谱检索数据集训练跨模态行人检索模型,模型就会更新参数,从而忘记行人检索的能力。而持续学习[102]研究如何从无限流数据中进行学习,在逐渐扩充新知识的同时保留旧知识。因此如何有效地结合跨模态检索和持续学习来提高跨模态检索模型的可塑性和稳定性是未来亟待解决的问题。

(2) 更高的检索效率

目前基于Transformer的预训练模型在多模态领域取得了极大的进展,仅需在跨模态检索任务上对预训练模型进行微调就能达到很好的效果,但是这些预训练模型参数量巨大,限制了其应用场景。目前跨模态哈希被用来提高跨模态检索效率,但是哈希码的稀疏性会导致模型精度的下降。因此如果能在预训练的过程中使用模型压缩在降低计算量的同时保持精度,将会为图文跨模态检索模型在实际应用中提供更多可能。

(3) 更完备的数据集

近年来跨模态检索模型越来越复杂,性能也逐渐提升。但是目前的数据集规模较小,类别不够丰富,不够完备的数据集已经成为限制模型性能进一步提升的瓶颈。比如常用的MS-COCO数据集[27]仅包含81个类别,其样本种类丰富度较日常生活所见还有很大差距,很难应用到实际场景中。因此,扩充数据集的样本数量,对数据集进行更加充分的标注,能够极大促进跨模态检索的实际应用与发展。

参考文献:

[1] SALTON G. Developments in automatic text retrieval[J].Science, 1991, 253: 974-980.

[2] DATTA R, JOSHI D, LI J, et al. Image retrieval: Ideas, influences, and trends of the new age[J]. ACM Computing Surveys (CSUR), 2008, 40 (2): 1-60.

[3] ASLANDOGAN Y A, YU C T. Techniques and systems for image and video retrieval[J]. IEEE Transactions on Knowledge and Data Engineering, 1999, 11(1): 56-63.

[4] GUDIVADA V N, RAGHAVAN V V. Content based image retrieval systems[J]. Computer, 1995, 28(9): 18-22.

[5] SNOEK C G, WORRING M. Concept-based video retrieval[J]. Foundations and Trends in Information Retrieval, 2009, 2(4): 215-322.

[6] SALVADOR A, HYNES N, AYTAR Y, et al. Learning cross-modal embeddings for cooking recipes and food images[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Honolulu, HI, USA: IEEE, 2017: 3020-3028.

[7] ZHANG Q, LEI Z, ZHANG Z, et al. Context-aware attention network for image-text retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle, WA, USA: IEEE, 2020: 3536-3545.

[8] DONG J, LI X, XU C, et al. Dual encoding for video retrieval by text[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021. DOI: 10.1109/TPAMI.2021.3059295.

[9] WANG J, HE Y, KANG C, et al. Image-text cross-modal retrieval via modality-specific feature learning[C]//Proceedings of the 5th ACM on International Conference on Multimedia Retrieval(ICMR). New York, NY, United States: ICMR, 2015: 347-354.

[10] ANDREW G, ARORA R, BILMES J, et al. Deep canonical correlation analysis[C]//Proceedings of International Conference on Machine Learning(ICML). Atlanta, GA, USA: ICML, 2013: 1247-1255.

[11] LIU J, XU C, LU H. Cross-media retrieval: State-of-the-art and open issues[J]. International Journal of Multimedia Intelligence and Security, 2010, 1(1): 33-52.

[12] XU C, TAO D, XU C. A survey on multi-view learning[EB/OL]. (2013-04-20)[2022-11-30]. https://arxiv. org/abs/1304.5634.

[13] WANG K, YIN Q, WANG W, et al. A comprehensive survey on cross modal retrieval[EB/OL]. (2016-07-21)[2022-11-30]. https://arxiv.org/abs/1607.06215.

[14] PENG Y, HUANG X, ZHAO Y. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges[J]. IEEE Transactions on Circuits and Systems for Video Technology(TCSVT), 2017, 28(9): 2372-2385.

[15] CHOROWSKI J K, BAHDANAU D, SERDYUK D, et al. Attentionbased models for speech recognition[J]. Advances in Neural Information Processing Systems, 2015. DOI:10.1016/0167-739X(94)90007-8.

[16] LEE K, CHEN X, HUA G, et al. Stacked cross attention for image-text matching[EB/OL]. (2018-07-23)[2022-11-30]. https://arxiv.org/abs/1803.08024v1.

[17] JI Z, WANG H, HAN J, et al. Saliency-guided attention network for image-sentence matching[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). Seoul, Korea (South): IEEE, 2019: 5754-5763.

[18] SONG Y, SOLEYMANI M. Polysemous visual-semantic embedding for cross-modal retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long Beach, CA, USA: IEEE, 2019: 1979-1988.

[19] YU F, TANG J, YIN W, et al. Ernievil: Knowledge enhanced visionlanguage representations through scene graphs[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, California, USA: AAAI, 2021: 3208-3216.

[20] AHMAD K. Slandail: A security system for language and image analysis-project No: 607691[J]. Available at SSRN 3060047, 2017. DOI:10.2139/ssrn.3060047.

[21] LU J, BATRA D, PARIKH D, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[J]. Advances in Neural Information Processing Systems(NIPS), 2019, 32: 13-23.

[22] LI X, YIN X, LI C, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks[C]//Proceedings of European Conference on Computer Vision(ECCV). [S.l.]: Springer, 2020: 121-137.

[23] LI W, GAO C, NIU G, et al. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning[C]// Proceedings of Meeting of the Association for Computational Linguistics. [S.l.]: Association for Computational Linguistics, 2020.

[24] SONG J, YANG Y, HUANG Z, et al. Multiple feature hashing for realtime large scale near-duplicate video retrieval[C]//Proceedings of the 19th ACM International Conference on Multimedia(ICMR). New York, United States: ICMR, 2011: 423-432.

[25] DING G, GUO Y, ZHOU J. Collective matrix factorization hashing for multimodal data[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA: IEEE,2014: 2075-2082.

[26] 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(SIGIR). New York, United States: ACM, 2014: 415-424.

[27] CHEN X, FANG H, LIN T Y, et al. Microsoft coco captions: Data collection and evaluation server[EB/OL].(2015-04-03)[2022-11-30]. https://arxiv.org/abs/1504.00325.

[28] ZHANG Y, JIN R, ZHOU Z H. Understanding bag-of-words model: A statistical framework[J]. International Journal of Machine Learning and Cybernetics, 2010, 1(1): 43-52.

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

[30] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60: 91-110.

[31] TOMASI C. Histograms of oriented gradients[J]. Computer Vision Sampler, 2012: 1-6.

[32] 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.

[33] ZHENG W, ZHOU X, ZOU C, et al. Facial expression recognition using kernel canonical correlation analysis (KCCA)[J]. IEEE Transactions on Neural Networks, 2006, 17(1): 233-238.

[34] KIM T K, WONG S F, CIPOLLA R. Tensor canonical correlation analysis for action classification[C]//Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Minneapolis, MN, USA: IEEE, 2007: 1-8.

[35] RASIWASIA N, MAHAJAN D, MAHADEVAN V, et al. Cluster canonical correlation analysis[C]//Proceedings of Artificial Intelligence and Statistics. New York, United States: ACM, 2014: 823-831.

[36] XU X, LIN K, GAO L, et al. Learning cross-modal common representations by private-shared subspaces separation[J]. IEEE Transactions on Cybernetics, 2020. DOI: 10.1109/TCYB.2020.3009004.

[37] KOLENDA T, HANSEN L K, LARSEN J, et al. Independent component analysis for understanding multimedia content [C]//Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing. Martigny, Switzerland: IEEE, 2002: 757-766.

[38] 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 MM). New York, United States: ACM, 2010: 251-260.

[39] PEREIRA J C, VASCONCELOS N. Cross-modal domain adaptation for text-based regularization of image semantics in image retrieval systems[J]. Computer Vision and Image Understanding, 2014, 124: 123- 135.

[40] ZHU X, HUANG Z, SHEN H T, et al. Linear cross-modal hashing for efficient multimedia search[C]//Proceedings of the 21st ACM International Conference on Multimedia (ICMR). New York, United States: ACM, 2013: 143-152.

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

[42] TENENHAUS A, TENENHAUS M. Regularized generalized canonical correlation analysis[J]. Psychometrika, 2011, 76(2): 257-284.

[43] CARROLL J D. Generalization of canonical correlation analysis to three or more sets of variables[C]//Proceedings of the 76th Annual Convention of the American Psychological Association. Washington D C, USA: [s.n.], 1968: 227-228.

[44] LOPEZ-PAZ D, SRA S, SMOLA A, et al. Randomized nonlinear component analysis[C]//Proceedings of International Conference on Machine Learning (ICML).[S.l.]: JMLR, 2014: 1359-1367.

[45] SUN T, CHEN S. Locality preserving CCA with applications to data visualization and pose estimation[J]. Image and Vision Computing, 2007, 25(5): 531-543.

[46] WANG K, HE R, WANG L, et al. Joint feature selection and subspace learning for cross-modal retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(10): 2010-2023.

[47] SUN T, CHEN S, YANG J, et al. Discriminative canonical correlation analysis with missing samples[C]//Proceedings of 2009 WRI World Congress on Computer Science and Information Engineering. Los Angeles, CA, USA: IEEE, 2009, 6: 95-99.

[48] XU G, LI X, ZHANG Z. Semantic consistency cross-modal retrieval with semi-supervised graph regularization[J]. IEEE Access, 2020, 8: 14278-14288.

[49] WANG G, JI H, KONG D, et al. Modality-dependent cross-modal retrieval based on graph regularization[J]. Mobile Information Systems, 2020. DOI:10.1155/2020/4164692.

[50] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[51] HE L, XU X, LU H, et al. Unsupervised cross-modal retrieval through adversarial learning[C]// Proceedings of 2017 IEEE International Conference on Multimedia and Expo(ICME). Hong Kong, China: IEEE, 2017: 1153-1158.

[52] ZHEN L, HU P, WANG X, et al. Deep supervised cross-modal retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 10394-10403.

[53] XU T, LIU X, HUANG Z, et al. Early-learning regularized contrastive learning for cross-modal retrieval with noisy labels [C]//Proceedings of the 30th ACM International Conference on Multimedia (ACM MM). New York, United States: ACM, 2022: 629-637.

[54] FENG F, WANG X, LI R. Cross-modal retrieval with correspondence autoencoder[C]//Proceedings of the 22nd ACM International Conference on Multimedia (ACM MM). New York, NY, United States: ACM, 2014: 7-16.

[55] ZHAO Y, WANG W, ZHANG H, et al. Learning homogeneous and heterogeneous co-occurrences for unsupervised cross-modal retrieval[C]//Proceedings of 2021 IEEE International Conference on Multimedia and Expo (ICME). Shenzhen, China: IEEE, 2021: 1-6.

[56] ZHENG Z, ZHENG L, GARRETT M, et al. Dual-path convolutional image-text embeddings with instance loss[J]. ACM

Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2020, 16(2): 1-23.
[57] CHEN W, LIU Y, BAKKER E M, et al. Integrating information theory and adversarial learning for cross-modal retrieval[J]. Pattern Recognition, 2021, 117: 107983.
[58] LI Z, LU W, BAO E, et al. Learning a semantic space by deep network for cross-media retrieval[C]//Proceedings of DMS. Hyatt Regency, Vancouver, Canada: DMS, 2015: 199-203.
[59] WEI Y, ZHAO Y, LU C, et al. Cross-modal retrieval with CNN visual features: A new baseline[J]. IEEE Transactions on Cybernetics, 2016, 47(2): 449-460.
[60] HU P, ZHEN L, PENG D, et al. Scalable deep multimodal learning for cross-modal retrieval[C]//Proceedings of the 42nd international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). New York, United States: ACM, 2019: 635-644.
[61] 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, 2013, 36(3): 521-535.
[62] LIU Y, WU J, QU L, et al. Self-supervised correlation learning for cross-modal retrieval[J]. IEEE Transactions on Multimedia, 2022. DOI: 10.1109/TMM.2022.3152086.
[63] GEIGLE G, PFEIFFER J, REIMERS N, et al. Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval[J]. Transactions of the Association for Computational Linguistics, 2022, 10: 503-521.
[64] TANG J, WANG K, SHAO L. Supervised matrix factorization hashing for cross-modal retrieval[J]. IEEE Transactions on Image Processing, 2016, 25(7): 3157-3166.
[65] JIANG Q Y, LI W J. Deep cross-modal hashing[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 3232-3240.
[66] SU S, ZHONG Z, ZHANG C. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 3027-3035.
[67] ZHANG D, WU X J, XU T, et al. Two-stage supervised discrete hashing for cross-modal retrieval[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022. DOI: 10.1109/TSMC.2021.3130939.
[68] ZHU L, SHEN J, XIE L, et al. Unsupervised visual hashing with semantic assistant for content-based image retrieval[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 29(2): 472-486.
[69] ZHANG J, PENG Y. Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval[J]. IEEE Transactions on Multimedia, 2019, 22(1): 174-187.
[70] ZHANG P F, LI Y, HUANG Z, et al. Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval [J]. IEEE Transactions on Multimedia, 2021, 24: 466-479.
[71] CHEN D, CHENG M, MIN C, et al. Unsupervised deep imputed hashing for partial cross-modal retrieval[C]//Proceedings of 2020 International Joint Conference on Neural Networks (IJCNN). Glasgow, UK: IEEE, 2020: 1-8.
[72] MENG M, WANG H, YU J, et al. Asymmetric supervised consistent and specific hashing for cross-modal retrieval[J]. IEEE Transactions on Image Processing, 2020, 30: 986-1000.
[73] LIANG M, CAO X, DU J, et al. Dual-pathway attention based supervised adversarial hashing for cross-modal retrieval[C]// Proceedings of 2021 IEEE International Conference on Big Data and Smart Computing (BigComp). Jeju Island, Korea (South): IEEE, 2021: 168-171.
[74] CHEN Z, ZHONG F, MIN G, et al. Supervised intra- and inter-modality similarity preserving hashing for cross-modal retrieval[J]. IEEE Access, 2018, 6: 27796-27808.
[75] LI C X, YAN T K, LUO X, et al. Supervised robust discrete multimodal hashing for cross-media retrieval[J]. IEEE Transactions on Multimedia, 2019, 21(11): 2863-2877.
[76] PLUMMER B A, WANG L, CERVANTES C M, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015: 2641-2649.
[77] 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 (CIVR). New York, United States:

IEEE, 2009: 1-9.

[78] KRAPAC J, ALLAN M, VERBEEK J, et al. Improving web image search results using query-relative classifiers[C]// Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR). San Francisco, CA, USA: IEEE, 2010: 1094-1101.

[79] GRUBINGER M, CLOUGH P, MÜLLER H, et al. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems[C]//Proceedings of International Workshop on Image. Genoa, Italy: LREC, 2006.

[80] ZHAI X, PENG Y, XIAO J. Learning cross-media joint representation with sparse and semi-supervised regularization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2013, 24(6): 965-978.

[81] WANG B, YANG Y, XU X, et al. Adversarial cross-modal retrieval[C]//Proceedings of the 25th ACM International Conference on Multimedia (ACM MM). New York, United States: ACM, 2017: 154-162.

[82] PENG Y, HUANG X, QI J. Cross-media shared representation by hierarchical learning with multiple deep networks[C]// Proceedings of IJCAI. New York, United States: IJCAI, 2016: 3846-3853.

[83] PENG Y, QI J, HUANG X, et al. CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network[J]. IEEE Transactions on Multimedia, 2017, 20(2): 405-420.

[84] XU X, TIAN J, LIN K, et al. Zero-shot cross-modal retrieval by assembling autoencoder and generative adversarial network [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2021, 17(S1): 1-17.

[85] QIAN S, XUE D, FANG Q, et al. Adaptive label-aware graph convolutional networks for cross-modal retrieval[J]. IEEE Transactions on Multimedia (TMM), 2021. DOI: 10.1109/TMM.2021.3101642.

[86] KUMAR S, UDUPA R. Learning hash functions for cross-view similarity search[C]//Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI). Shenzhen, China: IJCAI, 2011.

[87] SONG J, YANG Y, YANG Y, 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, United States: ACM, 2013: 785-796.

[88] BRONSTEIN M M, BRONSTEIN A M, MICHEL F, et al. Data fusion through cross-modality metric learning using similarity-sensitive hashing[C]//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, CA, USA: IEEE, 2010: 3594-3601.

[89] ZHANG D, LI W J. Large-scale supervised multimodal hashing with semantic correlation maximization[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, California, USA: AAAI, 2014.

[90] LIN Z, DING G, HU M, et al. Semantics-preserving hashing for cross view retrieval[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015: 3864-3872.

[91] WANG D, GAO X, WANG X, et al. Semantic topic multimodal hashing for cross-media retrieval[C]//Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence(IJCAI). Shanghai, China: IJCAI, 2015.

[92] YANG E, DENG C, LIU W, et al. Pairwise relationship guided deep hashing for cross-modal retrieval[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2017: 31.

[93] ZHANG X, LAI H, FENG J. Attention-aware deep adversarial hashing for cross-modal retrieval[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Computer Vision, 2018: 591-606.

[94] HUANG Y, WU Q, SONG C, et al. Learning semantic concepts and order for image and sentence matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE, 2018: 6163-6171.

[95] LI K, ZHANG Y, LI K, et al. Visual semantic reasoning for image-text matching[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 4654-4662.

[96] WEI X, ZHANG T, LI Y, et al. Multi-modality cross attention network for image and sentence matching[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 10941-10950.

[97] LIU C, MAO Z, ZHANG T, et al. Graph structured network for image text matching[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 10921-10930.

[98] DIAO H, ZHANG Y, MA L, et al. Similarity reasoning and filtration for image-text matching[C]//Proceedings of the AAAI

Conference on Artificial Intelligence. Palo Alto, California, USA: AAAI, 2021: 1218-1226.

[99] CHEN H, DING G, LIU X, et al. IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 12655-12663.

[100] CHEN J, HU H, WU H, et al. Learning the best pooling strategy for visual semantic embedding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Nashville, TN, USA: IEEE, 2021: 15789-15798.

[101] CHEN Y C, LI L, YU L, et al. Uniter: Universal image-text representation learning[C]//Proceedings of European Conference on Computer Vision (ECCV). Glasgow, UK: Computer Vision, 2020: 104-120.

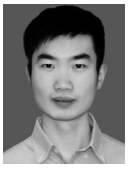
[102] HADSELL R, RAO D, RUSU A A, et al. Embracing change: Continual learning in deep neural networks[J]. Trends in Cognitive Sciences (TCS), 2020, 24(12):1028-1040.

作者简介:

张飞飞(1989-),**通信作者**,女,教授,硕士生导师,研究方向:多媒体计算、计算机视觉、模式识别、图像处理等,E-mail:feifeizhang@email.tjut.edu.cn。

马泽伟(1998-),男,硕士研究生,研究方向:跨模态检索。

周玲(1987-),女,博士,研究方向:模式识别、强化学习、调度优化,E-mail:zhouling_0922@163.com。



孟铃涛(1998-),男,硕士研究生,研究方向:跨模态检索。

(编辑:张黄群)