DOI:10.11918/202404027

实值无标签图文跨模态检索研究综述

张 力[1],陈 康[1],孙光辉[2]

(1.哈尔滨工业大学 计算学部,哈尔滨 150001;2.哈尔滨工业大学 航天学院,哈尔滨 150001)

摘 要:为研究面向无标签数据集基于实值特征的图像文本跨模态检索(以下简称跨模态检索)方法的发展现状和亟待解决的关键问题,对目前该领域的文献进行了分析与总结。跨模态检索是根据给定的一种模态查询,从另一种模态中检索出与查询相关的样本。首先,引入基于时间复杂度分类法,将现有跨模态检索方法分为基于特征方法和基于分数方法;其次,分别对以上两类方法的研究现状进行叙述,并针对两类方法现阶段存在的主要问题进行分析和讨论;然后,引入跨模态检索的两个主流数据集和常用评价指标,分别对两类方法在公开数据集上的性能进行比较与分析;最后,总结了跨模态检索领域亟待解决的关键问题。研究表明,现有跨模态检索方法尽管已经取得了显著进展,但仍有一些关键问题亟待解决,这些关键问题是未来跨模态检索领域的重要发展方向。

关键词:图像文本跨模态检索;多模态学习;实值特征;基于特征方法;基于分数方法

中图分类号:TP391.4 **文献标志码**:A **文章编号**:0367-6234(2024)09-0001-16

Review of unlabeled image-text cross-modal retrieval based on real-valued features

ZHANG Li[1], CHEN Kang[1], SUN Guanghui[2]

(1. Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China;
2. School of Astronautics, Harbin Institute of Technology, Harbin 150001, China)

Abstract: In order to investigate the current development status and key issues in the field of cross-modal retrieval based on real-valued features for unlabeled datasets (hereinafter referred to as cross-modal retrieval), this paper conducts an analysis and summary of the existing literatures. Cross-modal retrieval refers to the retrieval of samples from one modality that are relevant to a given query from another modality. Firstly, using a time complexity-based classification approach, existing cross-modal retrieval methods are categorized into feature-based methods and score-based methods. Secondly, the research status of these two categories of methods is described, and the main issues in the current stage for each category are analyzed and discussed. Furthermore, two mainstream datasets and commonly used evaluation metrics for cross-modal retrieval are introduced, and the performance of the two categories of methods on public datasets is compared and analyzed. Finally, key issues to be addressed in the field of cross-modal retrieval are summarized. The research indicates that although significant progress has been made in existing cross-modal retrieval methods, there are still key issues that urgently need to be addressed. These key issues represent important directions for future development in the field of cross-modal retrieval.

Keywords: image-text cross-modal retrieval; multimodal learning; real-valued feature; feature-based method; score-based method

互联网上存在着大量不同模态的数据[1-2],人们从海量数据中获取自己感兴趣的内容变得越来越困难。信息检索技术作为获取信息的重要手段,已经引起了人们的广泛关注[3]。信息检索技术按照处理数据类型的不同,可以分为单模态检索[4-7]和跨模态检索[8-12]。常见的单模态检索包括图像检索[13-15]、文本检索[16-18]和视频检索[19]等。典型的文本检索方法依赖一组关键词组成的查询,以此定位所需文档。如果某个文档包含更多查询关键词,则该文档相较于包含较少关键词的文档更具“相关性”。基于内容的图像检索提取待检索集的图像特征,并将这些特征存储于图像特征数据库。当给定

收稿日期:2024-04-11;**录用日期**:2024-05-29;**网络首发日期**:2024-07-22
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20240722.1304.22
基金项目:国家重点研发计划(2020AAA0106502);国家自然科学基金(62073105);机器人技术与系统国家重点实验室开放研究项目(SKLRS-2019-KF-14,SKLRS-202003D)
作者简介:张 力(1992—),男,助理研究员;孙光辉(1983—),男,教授,博士生导师
通信作者:张 力,zhangli92@hit.edu.cn

图像查询时,采用相同的方法提取其特征,并计算其与数据库中图像特征的相关度,从而实现检索。这是一种通用的单模态检索技术,可应用于文本检索、图像检索、音频检索和视频检索等。由于单模态检索方法只能处理特定类型的数据,这种局限性限制了其获取全面信息的能力。此外,在处理数据时,单模态检索方法也存在信息丢失和误差累积的问题,这些都会降低检索性能。因此,人们更需要能够跨越不同模态的检索技术[20],根据查询返回不同模态的结果,即跨模态检索技术,如图 1 所示。跨模态检索是根据给定的一种模态查询,从另一种模态中检索出与查询相关的样本。跨模态检索的主要问题是“模态鸿沟”[21-23],即不同模态的特征不一致,且位于不同的特征空间,因此度量不同模态特征之间的语义相似性成为一个具有挑战性的问题[24-25]。现有的跨模态检索方法可以分为两大类:一类是将两种模态数据通过特定模态的网络分支分别提取单模态特征,然后将两种单模态特征映射到公共语义空间中,采用不同的度量方法,通过测量特征之间的相似性,表达两种模态数据之间的语义相似性,从而弥合“模态鸿沟”[26](本文中称为基于特征方法);另一类是将融合特定模态分支提取的单模态特征送入多模态网络,多模态网络通过学习得出两种模态数据的相似性分数(本文中称为基于分数方法)。如今,研究人员针对图像、文本、音频和视频等模态间的跨模态检索进行了广泛研究(包括图像文本跨模态检索[27-32]、视频文本跨模态检索[33-35]和音频文本跨模态检索[36-38]等),但是图像文本跨模态检索无疑是其中受到最多关注的研究热点。其原因有以下几点:1)图像和文本是互联网中广泛存在的两种数据类型,大量的图像和文本数据集可供研究人员使用;2)现实生活中存在许多需要同时处理图像和文本的应用场景,图像文本跨模态检索技术可以显著改善用户体验和性能;3)图像文本跨模态检索融合了计算机视觉、自然语言处理和机器学习等多领域知识,跨学科的特性吸引了来自不同背景的研究人员。因此,本文也主要针对图像文本跨模态检索(以下简称跨模态检索)展开讨论。

虽然现有研究人员已经在上述两类方法(基于特征方法和基于分数方法)中产出了大量研究成果,但是目前跨模态检索方法在检索精度和检索效率上仍有提升的空间。因此,对现有成果进行分类、归纳和总结,对于跨模态检索领域的发展有重要意义。已经有一些综述对现有跨模态检索的研究进展进行了总结与归纳,例如:欧卫华等[39]给出了跨模态检索的核心思想、数据集和评价指标,指出了跨模态检索研究的未来趋势;徐文婉等[10]分别以实值和二值两种特征表示形式介绍了跨模态检索的研究进展,将跨模态检索方法分为公共空间学习、主题模型、深度学习、传统哈希和深度哈希 5 类;张飞飞等[12]将跨模态检索方法分为传统方法、基于深度学习方法和基于哈希表示方法,然后介绍了相关数据集,并在此基础上分别对每类方法进行分析。以上研究综述将基于实值特征方法和基于哈希表示方法视为跨模态检索任务的两类方法。但是,基于实值特征方法和基于哈希表示方法实际是跨模态检索领域的两类任务,其各自拥有不同的训练和测试数据集、不同的优化目标以及不同的评价指标。两类任务的跨模态检索方法也很难放在一起进行性能比较。杜锦丰等[8]从表示学习的角度,针对跨模态检索的各类方法,如基于统计相关分析、基于图推理和基于度量学习等方法展开综述;樊花等[9]从是否有监督的角度,针对哈希方法,对跨模态检索的线性模型和非线性模型展开综述;张博麟等[11]主要介绍了跨模态哈希的研究现状,重点阐述了生成对抗网络应用在基于哈希的跨模态检索方面的研究进展。但是以上研究综述没有区分有标签的跨模态检索数据集(如 Wikipedia[40]、NUS-WIDE[41]和 Pascal Sentence[42]等)和无标签的跨模态检索数据集(如 MS-COCO(Microsoft common objects in context)[43]和 Flickr30K[44]),如图 1 所示。面向有标签数据集的跨模态检索方法(一般是指跨模态哈希检索),根据网络的前向过程中是否使用标签,可以进一步分为有监督方法和无监督方法。面向无标签数据集的跨模态检索方法易与上述两类面向有标签数据集的方法混淆。

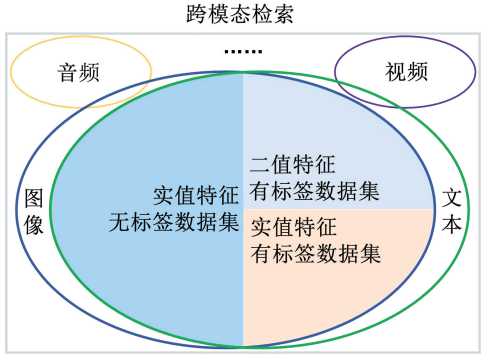


图 1　跨模态检索的分类

Fig. 1　Classification of cross-modal retrieval

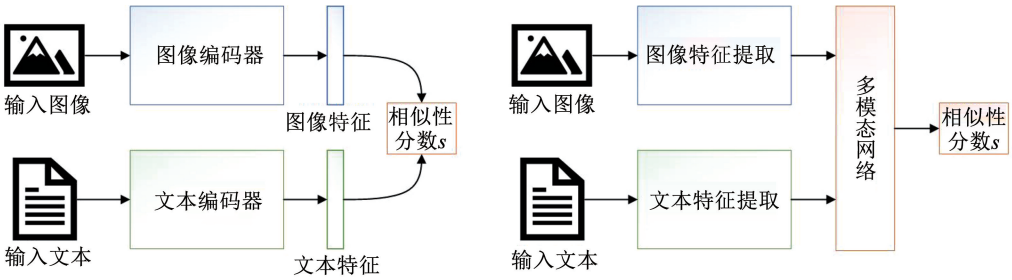
因此,本文主要针对面向无标签数据集基于实值特征的图像文本跨模态检索(以下简称跨模态检索)方法展开综述,将现有跨模态检索方法按照测试和推理的时间复杂度分为基于特征方法和基于分数方法分别展开论述。在此基础上,对跨模态检索

的两个主流数据集和常用的评价指标进行介绍，然后分别对基于特征方法和基于分数方法在公开数据集上的性能进行比较与分析，最终引出跨模态检索领域亟待解决的问题。

1　基于时间复杂度的跨模态检索分类法

按照测试和推理阶段的时间复杂度，现有的跨模态检索方法可大致分为两类：基于特征方法和基于分数方法。两类方法的训练过程见图 2。基于特征方法利用图像编码器和文本编码器分别提取图像特征和文本特征，将图像特征和文本特征映射到公共语义空间，采用特定的度量方法衡量特征之间的相似性；基于分数方法则将提取到的图像特征和文本特征一起送入多模态网络中预测图像-文本对的语义相似性。

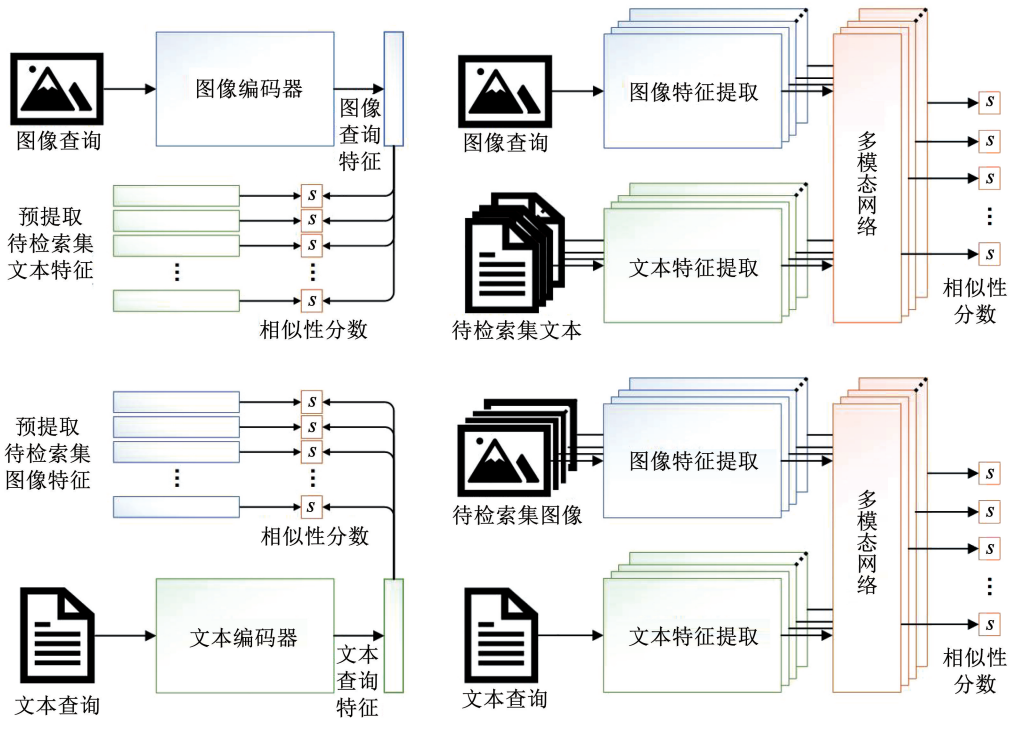
在测试和推理阶段，基于特征方法可以在预提取待检索集特征的情况下，仅将单个查询编码为全局特征，通过比较查询全局特征和待检索集全局特征，可以很快找到与查询最相关的待检索样本；而基于分数方法，需要将查询与待检索集中所有样本逐一配对，送入多模态网络计算相似性分数，通过比较所有样本对的相似性分数，找到与查询最相关的待检索样本。两类方法的测试和推理过程见图 3。



(a) 基于特征方法　　(b) 基于分数方法

图 2　两类图像文本跨模态检索方法的训练过程

Fig. 2　Training processes of two types of cross-modal retrieval methods for image-text retrieval



(a) 基于特征方法　　(b) 基于分数方法

图 3　两类图像文本跨模态检索方法的推理过程

Fig. 3　Inference processes of two types of cross-modal retrieval methods for image-text retrieval

值得注意的是,图 2 和图 3 的结构也兼容其他模态的跨模态检索,如视频文本跨模态检索中的 HGR[33]、TMMGT-GLA[34]和音频文本跨模态检索中的 SFA[36]、3CMLF[37]均采用基于特征方法的思想进行训练、测试与推理,视频文本跨模态检索中的 HCGC[35]和音频文本跨模态检索中的 LASO[38]均采用基于分数方法的思想进行训练、测试与推理。

表 1 和表 3 是近年来基于特征方法和基于分数方法在 MS-COCO 数据集 1K 和 5K 设置的性能比较情况。由于图像与文本之间缺乏交互,同一时期基于特征方法的性能普遍弱于基于分数方法。此外,由于基于 Transformer 的视觉语言预训练(vision language pre-training,VLP)模型的迅速发展,基于分数的跨模态检索方法在图像和文本的深度交互下取得了优异性能。

图 4 展示了典型的基于特征方法和基于分数方法在 Flickr30K 测试集上的测试时间。基于特征方法有 VSE++[45](41.8 s)、GPO[46](129.5 s)、M2EF[47](130.1 s)、VSRN[31](667.6 s+32.6 s)、CAMERA[29](667.6 s+13.5 s)和 DCPG[48](667.6 s+36.9 s),基于分数方法有 SCAN[49](667.6 s+210.1 s)、GSMN[50](667.6 s+235.5 s)和 ALBEF[51](2 185.9 s)。所有测试时间均由具有单个 NVIDIA RTX 2080 Ti GPU(11G)的本地工作站获得,其中斜线部分(667.6 s)表示图像分支采用自底向上的注意力(bottom-up attention,BUA)模型进行特征提取所需时间。

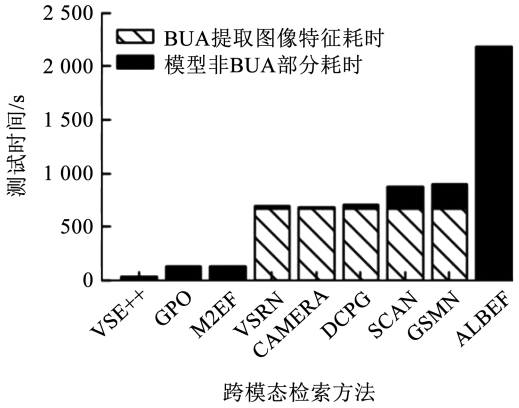


图 4　两类方法在 Flickr30K 测试集上的测试时间

Fig. 4　Testing time of two types of methods on the Flickr30K test set

由以上分析可知,基于特征方法由于缺乏图像文本交互,使其无法衡量不同模态局部片段之间的相似性,无法学习图像区域和文本单词之间的对应关系,这些都会对最终度量图像和文本的语义相似性造成不良影响,进而影响检索精度。而基于分数方法在推理过程中需要将单个查询与待检索集中所有样本分别两两配对,再输入多模态网络计算语义相似性,导致其时间开销巨大,难以应用于真实场景。

2　研究现状概述

基于特征方法和基于分数方法的时间复杂度分析见图 5。在测试阶段,对于包含 n 个图像-文本对的测试集,基于特征方法的时间复杂度为 $O(2n)$

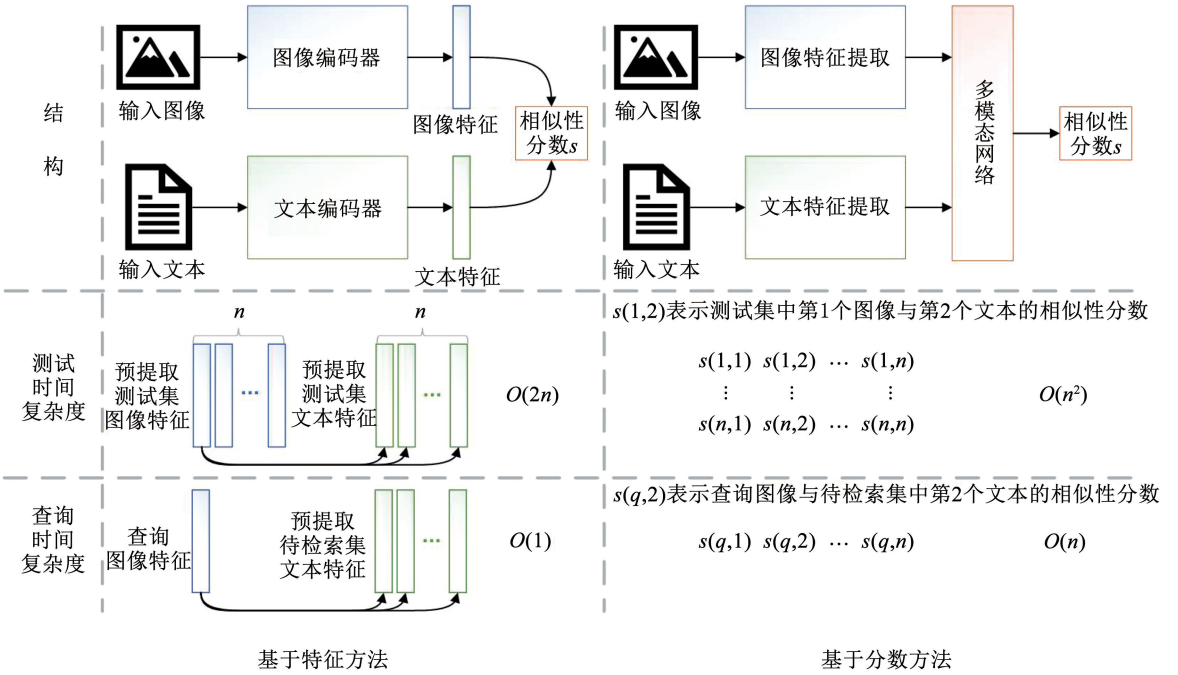


图 5　两类图像文本跨模态检索方法的时间复杂度分析

Fig. 5　Time complexity analysis of two types of image-text cross-modal retrieval methods

（假设图像或文本通过相应分支的时间复杂度为 $O(1)$，忽略图像和文本特征之间计算相似性的时间复杂度）。给定 1 个查询，待检索集包含 n 个样本，在预提取待检索集特征的条件下，查询的时间复杂度为 $O(1)$。而同样在测试阶段，对于包含 n 个图像–文本对的测试集，$s(i,j)$ $(i,j\in[1,n])$ 表示第 i 个图像和第 j 个文本的相似性，基于分数方法的时间复杂度为 $O(n^2)$（假设图像和文本通过网络的时间复杂度为 $O(1)$）。给定 1 个查询 q 和包含 n 个样本的待检索集，$s(q,j)$ $(j\in[1,n])$ 表示查询 q 与第 j 个跨模态实例的相似性，查询的时间复杂度为 $O(n)$。人们对这两类跨模态检索方法进行了大量的研究。本节将分别对基于特征方法和基于分数方法的研究现状、现阶段存在的主要问题进行叙述、分析和讨论。

2.1　基于特征的跨模态检索方法

早期基于特征的跨模态检索方法采用图像分类网络编码图像，因此图像编码器的输出是可用于分类的图像全局特征向量。例如，Kiros 等[52] 提出结构内容神经语言模型（structure content-neural language model，SC-NLM），利用深度卷积神经网络（convolutional neural network，CNN）和长短期记忆（long short term memory，LSTM）网络作为编码器，学习联合图像文本嵌入空间。SC-NLM 的解码器以嵌入空间中的图像特征为内容，结合文本结构信息，以序列方式将图像特征解码为句子，辅助跨模态检索任务的训练。Wang 等[53] 提出深度结构保持图像文本嵌入（deep structure-preserving image-text embeddings，DSPE）模型，基于对比学习引入模态内邻域保持约束，以优化跨模态检索的目标函数。Ren 等[54] 提出高斯视觉语义嵌入（gaussian visual semantic embedding，GVSE）模型，利用视觉信息将文本概念建模为语义空间中的高斯分布，从而更好地捕捉每个文本概念的不确定性，并能够更好地对包含和交叉等概念进行几何解释。Faghri 等[45] 提出增强视觉语义嵌入（improving visual semantic embedding，VSE++）框架，将难负样本挖掘引入跨模态检索任务的三元排名损失之中，获得显著性能增益。

一些工作在上述方法的基础上引入额外的模块或任务，以促进跨模态方法的检索性能。Wang 等[55] 提出对抗跨模态检索（adversarial cross-modal retrieval，ACMR）方法，在对抗学习的基础上寻找有效的公共子空间。ACMR 包含两个相互作用的模块：一个是特征映射器，试图在公共子空间中生成一种模态不变的表示，并混淆另一个模块；另一个是模态分类器，试图根据生成的表示区分不同的模态。Engilberge 等[47] 提出深度语义视觉嵌入（deep semantic visual embedding，DSVE）模型，利用双流神经网络将图像和文本分别映射到相同的共享欧几里德空间，用于捕获有用的语义关系。DSVE 模型生成的多模态特征还可以用于视觉定位任务并取得当时最先进的结果，扩展了语义视觉架构在视觉定位任务方面的应用。Gu 等[56] 提出生成式跨模态网络（generative cross-modal network，GXN），将图像描述任务和文本生成图像任务整合到公共空间图像文本全局特征表示学习中，以促进跨模态检索的性能。Zhang 等[57] 提出基于特征分离与重构的跨模态检索方法，引入特征分离解决不同模态之间的信息不对称问题，并引入图像和文本重构任务，通过多任务联合学习提高跨模态检索任务的性能。

然而，上述基于特征方法采用的图像分类网络倾向于提取图像中最显著目标的信息，这使得图像全局特征向量中其他目标的语义信息被忽略，进而导致跨模态检索中的错误匹配。2018 年，Anderson 等[58] 利用 Visual Genome 数据集[59] 上训练的 Faster R-CNN[60] 为输入图像提取一组显著图像区域，每个区域由池化后的卷积特征向量表示。此后，不少基于特征方法开始采用 BUA 作为图像编码器，用一组区域特征表示输入图像，获得了比用分类网络作为图像编码器更好的性能。例如，Li 等[27] 提出基于动作感知记忆增强嵌入（action-aware memory-enhanced embedding，AME）的跨模态检索方法，集成动作预测与动作感知记忆库集，以动作相似的文本特征丰富当前的图像和文本特征，将动作信息增强的图像和文本特征映射到共享嵌入空间；Yan 等[48] 提出基于离散连续策略梯度（discrete continuous policy gradient，DCPG）的跨模态检索方法，利用离散–连续策略梯度分别为图像区域特征和文本单词特征生成注意力权重，应用注意力权重对图像区域特征或文本单词特征进行加权融合，最后采用度量学习损失、离散策略梯度损失和连续策略损失训练整个模型；Zhu 等[61] 提出即插即用的外部空间注意力聚合（external space attention aggregation，ESA）模块，在视觉语言嵌入框架的基础上，通过引入外部记忆与局部特征执行注意力机制得到向量通道级别的权重矩阵，局部特征利用权重矩阵计算全局特征以用于跨模态检索；梁彦鹏等[62] 提出嵌入共识知识的因果图文检索方法（causal image-text retrieval methodology with embedded consensus knowledge，CITR），将因果干预引入视觉特征提取模块，通过因果关系替换相关关系，学习常识因果视觉特征，并与原始视觉特征进行级联得到最终的视觉特征表示。

一部分工作在此基础上引入图卷积网络(graph convolutional network,GCN),推理局部特征之间的语义关系。Li 等[31]提出视觉语义推理网络(visual semantic reasoning network,VSRN),通过捕获场景图中关键目标和语义概念生成视觉表示,然后利用GCN[63]进行区域关系推理,利用门控机制和记忆机制进行全局语义推理,最后生成图像特征表示。Zhang 等[64]提出跨模态多关系感知推理网络(cross-modal multi-relationship aware reasoning network,CMRN),用于提取几何位置关系和语义交互关系并学习图像区域之间的相关性。CMRN 将图像处理为图模型,通过引入空间关系编码器,利用具有注意力机制的 GCN 对图模型进行推理。

然而,上述基于特征方法存在的问题是:1)图像和文本模态往往被独立编码,然后在特征空间中进行匹配。这种方式忽略了不同模态之间的潜在关联,导致模态间交互不足,进而影响检索精度;2)现有方法在局部特征聚合时存在信息损失(图像区域特征聚合为图像全局特征可能丢失局部特征中的重要细节,文本单词特征聚合为文本全局特征可能难以建模长期依赖关系),导致全局特征不够鲁棒,进而影响检索精度。

2.2 基于分数的跨模态检索方法

基于分数的跨模态检索方法主要学习输入图像-文本对的相似性分数。Karpathy 等[65]基于区域卷积神经网络(region convolutional neural network,RCNN)[66]和双向递归神经网络(bidirectional recurrent neural network,Bi-RNN)[67]提出深度视觉语义对齐(deep visual semantic alignments,DVSA)模型,利用图像和对应的图像描述学习语言和视觉数据之间的跨模态对应关系;Huang 等[68]提出语义概念和顺序(semantic concepts and order,SCO)模型,通过使用多区域多标签 CNN 学习目标、属性和动作等语义概念,并按照正确的语义顺序组织语义概念以改进图像表示。

作为基于分数方法的典型代表,Lee 等[49]提出堆叠交叉注意力网络(stacked cross attention network,SCAN),通过计算图像中显著区域和相应句子中单词之间的潜在语义对齐捕捉视觉和语言之间的细粒度相互作用,从而推断图像文本的相似性分数。基于 SCAN,研究人员提出了许多改进方法。例如,Chen 等[30]提出带有语义一致性的跨模态检索方法(cross-modal retrieval with semantic consistency,CMR-SC),通过引入语义一致性以联合学习不同的嵌入空间。CMR-SC 构建了两个不同的嵌入空间,即基于图像嵌入空间和基于文本嵌入空间,并在公共排名目标函数中加入语义一致性约束,同时学习两个嵌入空间并相互约束,从而提升跨模态检索的性能。Huang 等[69]提出对齐跨模态记忆(aligned cross-modal memory,ACMM)模型以研究跨模态检索任务中小样本内容带来的挑战。ACMM 包含对齐的记忆控制器网络,用于产生两组语义可比的接口向量。记忆项持久地记忆跨模态共享语义表示,接口向量通过与其进行交互以增强小样本内容表示。Wang 等[70]提出跨模态自适应消息传递(cross-modal adaptive message passing,CAMP)模型,由消息聚合模块和门控融合模块组成。消息聚合模块将对应于每个单词的显著视觉信息聚合为从视觉模态传递到文本模态的消息,同时将对应于每个区域的显著文本信息聚合为从文本模态传递到视觉模态的消息。门控融合模块在两个模态特征相互融合的过程中自适应地控制融合程度,从而消除融合过程中不匹配区域单词对的影响。

然而,上述方法没有考虑到局部特征之间的关系,会导致模型无法区分具有相同局部特征但不同关系的复杂场景(如“人骑马”和“人牵马”)。在此基础上,部分工作引入 GCN,用于推理局部特征之间的语义关系。Wang 等[71]提出场景图匹配(scene graph matching,SGM)模型,分别从图像和文本中提取目标和关系,生成视觉场景图(visual scene graph,VSG)和文本场景图(textual scene graph,TSG)。SGM 模型包含的两个图编码器将 VSG 和 TSG 编码为视觉特征图和文本特征图。之后,SGM 在每个特征图中学习目标级和关系级的特征,最终在目标和关系层次上分别匹配视觉特征图和文本特征图。Liu 等[50]提出图结构化匹配网络(graph structured matching network,GSMN),引入视觉图模型和文本图模型对图像文本的视觉一致性和文本一致性进行建模,结合节点级匹配和结构级匹配衡量图像和文本之间的相似性。Diao 等[72]提出相似性图推理和注意力过滤(similarity graph reasoning and attention filtration,SGRAF)网络,通过相似性图推理模块(similarity graph reasoning,SGR)建立包含全局对齐和局部对齐的图模型,实现全局和局部对齐之间的信息传递,获取更全面的交互以促进相似性预测;通过相似性注意力过滤模块(similarity attention filtering,SAF)增强重要的细粒度对齐,并抑制无效对齐。刘长红等[73]提出基于语义关系图的跨模态张量融合网络(cross-modal tensor fusion network based on semantic relation graph,CMTFN-SRG),利用图卷积和双向门控循环单元(gated recurrent unit,GRU)分别学习模态内局部特征之间的关系,利用

张量融合模块学习两种不同局部特征之间的语义关联，进而捕获全局语义相关性。Ge 等[74]提出跨模态语义增强交互（cross-modal semantic enhanced interaction，CMSEI）方法，利用两个关系感知 GCN 分别整合显著区域的空间关系和语义关系，利用预训练基于 Transformer 的双向编码器（bidirectional encoder representations from transformers，BERT）获得的高级语义信息对显著区域进行两种关系的增强，最终借助跨模态对齐度量图像和文本之间的语义相似性。

上述基于图卷积的方法有效地建模了局部特征之间的关系，使图像文本匹配不仅局限于局部特征匹配，还对齐了不同模态之间的复杂语义概念。然而，基于图卷积的方法通常将局部特征视为同等重要，构建图模型并执行关系推理，但是在图像场景和图像描述中，人们往往侧重显著目标而忽略次要部分。因此，部分工作利用全局特征指导局部特征执行细粒度相似性度量。Chen 等[75]提出自适应置信度匹配网络（adaptive confidence matching network，ACMNet），用于处理细粒度区域单词匹配方法中存在的相似性偏差问题。在此基础上，ACMNet 利用全局文本（图像）信息预测局部相似性的置信度分数，用于加权区域（单词）与整个文本（图像）的语义相关性。Zhang 等[28]提出上下文感知注意力网络（context-aware attention network，CAAN），通过聚合全局上下文，有选择地关注关键局部片段（区域和单词），同时利用全局模态间对齐和模态内关联发现潜在的语义关系。Liu 等[76]提出双向矫正注意力网络（bi-directional correct attention network，BCAN），利用全局矫正单元和局部矫正单元分别对局部特征对应的上下文特征进行矫正，得到图像隐空间和文本隐空间中局部特征与矫正后的上下文特征的相似性，最终预测图像文本的语义相似性。Zhang 等[77]提出增强语义相似性学习（enhanced semantic similarity learning，ESL），构造不同层级的测量单元用于表示局部特征，然后在对应的测量单元中学习局部特征一致性，最终得到细粒度相似性矩阵，用于训练跨模态检索模型。杨晓宇等[78]提出分层聚合共享网络（hierarchical aggregation sharing network，HAS），将 BUA 提取的图像区域特征和 BERT 提取的文本单词特征分别送入权值共享的 Transformer 编码器中，通过分层结构获得基本语法信息和高级语义信息，进而聚合为全局特征，用于跨模态检索模型的训练。魏钰琦等[79]提出跨模态信息交互推理网络（cross-modal information interaction reasoning network，CMIIRN），包含自适应交叉注意力模块和关系推理模块。自适应交叉注意力模块用于交互关注和减弱冗余信息影响；关系推理模块用于迭代地加入融合后的增强信息，逐步推理全局语义信息。Wu 等[80]提出双视角语义推理（dual view semantic inference，DVSI）网络，用于在网络中同时利用局部语义匹配信息和全局语义匹配信息。对于局部视角，DVSI 提出区域增强模块以挖掘图像中不同区域的优先级；对于全局视角，DVSI 利用图像和句子的整体语义进行全局语义匹配，以避免全局语义漂移。最后，统一两个视角以得到图像文本的相似性分数。

除了引入全局特征指导局部特征的细粒度相似性度量外，还有部分工作引入知识蒸馏和强化学习等技术以强化细粒度对齐。Chen 等[81]提出带有循环注意力记忆的迭代匹配（iterative matching with recurrent attention memory，IMRAM）网络，通过引入迭代匹配以逐步探索图像和文本之间的细粒度对应关系；还引入了记忆蒸馏单元以更新查询特征，进而用于细化对齐知识。Zhang 等[82]提出基于知识蒸馏和隐空间语义监督的跨模态检索模型（latent space semantic supervision model based on knowledge distillation，L3S-KD），将来自目标检测模型的目标分类器和属性分类器引入图像隐空间，并通过知识蒸馏将目标检测模型中的语义知识迁移到图像隐空间；然后将目标分类器和属性分类器引入文本隐空间，用于对齐文本单词特征和对应的单词上下文特征。Cai 等[83]引入部分查询问题，提出基于弱监督强化学习的交互式检索框架“询问和确认（ask-and-confirm）”。该框架中的代理首先根据初始文本查询，从数据集中检索一组相关的候选图像供用户确认。根据用户确认，代理缩小候选图像的范围，最终收集足够的信息以定位目标图像。

上述方法均在图像-文本对中探索图像区域与所有单词的细粒度相似性，以及单词与所有图像区域之间的细粒度相似性。这些方法固然可以建模“一对多”关系，用于衡量图像与文本之间的相似性，但是缺乏多模态局部特征“多对多”的建模能力，以及通过预训练从海量图像-文本对中学习多模态特征的能力。随着基于 Transformer[84] 的 VLP[85-88] 技术的迅速发展，基于分数的跨模态检索也有了显著进展。与上述非 VLP 方法相比，基于分数的 VLP 跨模态检索方法的性能有了很大提高。例如，Li 等[89]提出目标语义对齐预训练（object-semantics aligned pre-training，Oscar）方法，将图像中检测到的目标标签作为锚点，从两个视角执行预训练任务。在字典视角，Oscar 连接文本单词特征与锚点特征，再与图像区域特征一起送入 Transformer，利

用自注意力机制学习图像文本语义对齐,利用掩码令牌损失(masked token loss,MTL)[90]模型进行训练;在模态视角,Oscar 连接图像区域特征与锚点特征,再与文本单词特征一起送入 Transformer,利用自注意力机制学习图像文本语义对齐,利用对比损失模型进行训练。Park 等[91]提出语义对齐模块(semantic alignments module,SAM),将 VLP Transformer 编码器输出的图像区域特征和文本单词特征视为节点,通过计算不同模态节点之间的注意力系数,对所有节点进行更新,然后送入二值分类器,用于度量图像和文本的语义相似性。陈曦等[92]提出基于预训练模型和编码器的图文跨模态检索(cross-modal retrieval based on pre-trained models and encoders,PTME)方法,利用双编码器实现粗略召回,利用融合编码器实现精准排序;并提出了基于多路 Transformer 的双编码器和融合编码器,实现图文之间高质量语义对齐,提升了检索性能。

Zhang 等[85]研究了视觉语言模型中视觉表示(visual representations in vision-language models,VinVL)的改进方法,通过丰富视觉目标和属性类别、扩大模型大小和在更大规模的目标检测数据集上进行预训练等手段,开发一种改进的目标检测模型。而后将新的目标检测模型生成的区域特征输入 Oscar 模型进行预训练,并在广泛的下游任务中进行微调。Liu 等[32]提出跨模态语义重要性一致性网络(cross-modal semantic importance consistency,CSIC),在图像编码器中引入图像区域特征和位置信息,在文本编码器中引入文本 token 和位置信息,共同送入多模态 Transformer 中进行训练,并分别在图像和文本中融合模态内注意力和模态间注意力。

现有基于分数方法凭借图像和文本局部特征的深度交互,在跨模态检索任务取得了优异性能。然而,基于分数方法存在测试和推理时间复杂度高的问题,导致其难以在真实场景的大规模数据集上应用。

3 数据集、评价指标与性能比较

由于数据集直接影响模型训练和性能评估,因此数据集的选择对于跨模态检索任务至关重要。目前,跨模态检索主流数据集是 MS-COCO[43]和 Flickr30K[44]。这两个数据集均包含图像和文本两种模态的数据,具有代表性和丰富性,有助于确保算法的泛化能力和有效性。同时,评价指标可以客观衡量模型的性能,并帮助研究者进行方法之间的公平比较,因此评价指标的选择也至关重要。接下来,将对跨模态检索的两个主流数据集和常用评价指标进行详细介绍,然后分别对基于特征方法和基于分数方法在公开数据集上的性能进行比较与分析。

3.1 跨模态检索常用的数据集

MS-COCO 数据集由微软研究院创建,包含超过百万个图像,这些图像涵盖了多种场景,包括人类和动物的行为、室内外的物体等,具有很高的代表性和多样性。所有图像都经过了精细标注,即每个图像都包含多个物体的位置信息和与之相关的文本描述,这些标注常用于图像理解、图像描述和跨模态检索等任务。在 MS-COCO 数据集上可以进行多种任务的研究,包括目标检测、物体识别、场景理解和图像描述生成等,这也使得该数据集成为广泛用于计算机视觉、自然语言处理和多模态学习领域的基准数据集之一。

应用于跨模态检索任务中的 MS-COCO 数据集包含 123 287 个图像,每个图像对应 5 个描述图像的文本。根据 VSE++[45]和 SCAN[49]的设置,MS-COCO 被划分为训练集、验证集和测试集 3 部分。其中,训练集包含 113 287 个图像及其对应文本;验证集和测试集分别包含 5 000 个图像及其对应文本。此外,根据测试集划分方式的不同,MS-COCO 数据集的测试可以进一步细分为 1K 和 5K 设置。1K 设置表示将测试集划分为 5 个部分,每个部分包含 1 000 个图像和对应文本,用于跨模态检索测试,然后计算 5 部分测试结果的平均值;5K 设置表示采用测试集中的 5 000 个图像和对应文本进行跨模态检索测试。

Flickr30K 是一个用于跨模态检索的常用数据集,包含来自 Flickr 图片分享平台的31 000 个图像,这些图像涵盖了各种场景,包括人类活动和自然风光等,具有很高的代表性和多样性。每个图像对应 5 个人工标注的文本描述,涵盖了图像中的场景、对象和人物等信息。这些描述由众包方式生成,以确保数据的多样性和丰富性。Flickr30K 数据集中图像与文本描述之间的对应用于探索图像和文本之间的语义联系,为研究者提供了理想的数据支持。因此,Flickr30K 数据集极大地促进了多模态学习研究的发展。

根据VSE++[45]和 DVSA[65]的设置,跨模态检索任务中采用的 Flickr30K 数据集同样被划分为训练集、验证集和测试集 3 部分。其中,训练集包含 29 000 个图像及其对应文本;验证集和测试集分别包含 1 000 个图像及其对应文本。

3.2 评价指标

跨模态检索任务一般采用$R_K(K=1,5,10)$(表示测试集中所有查询的前 K 个检索结果包含相关

项的百分比）作为评价指标，用于衡量检索系统的性能。对于给定查询的检索结果列表，如果前 K 个检索结果中包含与查询相关的项，则记为 1；如果不包含，则记为 0。在跨模态检索中，假设测试集包含 1 000 个图像和 5 000 个文本，那么与任意图像相关的文本有 5 个，与任意文本相关的图像仅有 1 个。在执行过程中，当选择所有图像分别作为查询时，所有文本作为待检索集；当选择所有文本分别作为查询时，所有图像作为待检索集。

此外，为了综合判断模型的检索性能，R_s 被引入作为评价指标。R_s 表示图像检索文本和文本检索图像的所有 R_K（$K=1,5,10$）之和，具体计算公式如下：

$$R_s = R_{i1} + R_{i5} + R_{i10} + R_{t1} + R_{t5} + R_{t10}$$

式中：R_{i1}、R_{i5} 和 R_{i10} 分别表示图像检索文本的 R_1、R_5 和 R_{10}，R_{t1}、R_{t5} 和 R_{t10} 分别表示文本检索图像的 R_1、R_5 和 R_{10}。

3.3　性能比较

近年来，基于特征方法在 MS-COCO 数据集 1K 和 5K 设置下的性能比较见表 1，在 Flickr30K 数据集上的性能比较见表 2。

表 1　近年来基于特征方法在 MS-COCO 数据集上的性能比较

Tab. 1　Performance comparison of feature-based methods on the MS-COCO dataset in recent years

年份	方法	MS-COCO 1K							MS-COCO 5K						
		图像检索文本			文本检索图像			R_s	图像检索文本			文本检索图像			R_s
		R_1	R_5	R_{10}	R_1	R_5	R_{10}		R_1	R_5	R_{10}	R_1	R_5	R_{10}	
2014	SC-NLM[52]														
2016	DSPE[53]	50.1	79.7	89.2	39.6	75.2	86.9	420.7							
2018	VSE++[45]	64.6	90.0	95.7	52.0	84.3	92.0	478.6	41.3	71.1	81.2	30.3	59.4	72.4	355.7
2018	DSVE[47]	69.8	91.9	96.6	55.9	86.9	94.0	495.1							
2018	GXN[56]	68.5		97.9	56.6		94.5		42.0		84.7	31.7		74.6	
2019	VSRN[31]	76.2	94.8	98.2	62.8	89.7	95.1	516.8	53.0	81.1	89.4	40.5	70.6	81.1	415.7
2021	DCPG[48]	84.0	95.8	97.8	63.9	88.9	95.6	526.0	68.7	88.7	93.0	46.2	77.8	85.5	459.9
2022	CMRN[64]	73.9	93.9	97.9	60.4	88.5	94.0	508.6							
2022	AME[27]	79.4	96.7	98.9	65.4	91.2	96.1	527.7	59.9	85.2	92.3	43.6	72.6	82.7	436.3
2023	ESA[61]	81.0	96.9	98.9	66.4	92.2	96.5	531.9	61.1	86.6	92.9	43.9	74.1	84.4	443.0
2024	CITR[62]	78.6	96.4	98.9	62.8	90.4	96.3	523.4	55.3	84.3	91.7	42.4	71.7	81.4	426.8

表 2　近年来基于特征方法在 Flickr30K 数据集上的性能比较

Tab. 2　Performance comparison of feature-based methods on the Flickr30K dataset in recent years

年份	方法	图像检索文本			文本检索图像			R_s
		R_1	R_5	R_{10}	R_1	R_5	R_{10}	
2014	SC-NLM[52]	23.0	50.7	62.9	16.8	42.0	56.5	251.9
2016	DSPE[53]	40.3	68.9	79.9	29.7	60.1	72.1	351.0
2018	VSE++[45]	52.9	80.5	87.2	39.6	70.1	79.5	409.8
2018	DSVE[47]	46.5	72.0	82.2	34.9	62.4	73.5	371.5
2018	GXN[56]	56.8		89.6	41.5		80.1	
2019	VSRN[31]	71.3	90.6	96.0	54.7	81.8	88.2	482.6
2021	DCPG[48]	82.8	95.9	97.9	62.2	89.3	93.8	521.9
2022	CMRN[64]	70.8	91.5	95.4	55.2	81.8	88.1	482.8
2022	AME[27]	81.9	95.9	98.5	64.6	88.7	93.2	522.8
2023	ESA[61]	84.6	96.6	98.6	66.3	88.8	93.1	528.0
2024	CITR[62]							

由表 1 和表 2 可以看出,近年来基于特征方法在 MS-COCO 数据集和 Flickr30K 数据集上的性能都有了稳步提升,且精度相差不大。在包含 1 000 个图像和 5 000 个文本的同样规模的测试集中(Flickr30K 测试集和 MS-COCO 1K 设置,即每个图像查询需要从 5 000 个候选文本中,至少检索出 5 个对应文本中的 1 个;每个文本查询需要从 1 000 个候选图像中,检索出对应的 1 个图像),对于图像检索文本,R_1的最好结果达到了 84.6,R_5的最好结果达到了 96.9,R_{10}的最好结果达到了 98.9;对于文本检索图像,R_1的最好结果达到了 66.4,R_5的最好结果达到了 92.2,R_{10}的最好结果达到了 96.5;R_s的最好结果达到了 531.9,体现出了基于特征方法在跨模态检索任务方面的鲁棒性和稳定性。而在 MS-COCO 5K 设置中,由于测试集包含 5 000 个图像和 25 000 个文本,也就是说,每个图像查询需要从 25 000 个候选文本中,至少检索出 5 个对应文本中的 1 个;每个文本查询需要从 5 000 个候选图像中,检索出对应的 1 个图像,所以基于特征方法在 MS-COCO 5K 设置下的性能都低于其在 Flickr30K 测试集和 MS-COCO 1K 设置下的性能。此外,对于任意基于特征方法,其在图像检索文本子任务的性能都优于其在文本检索图像子任务的性能,这是因为在图像检索文本子任务中,1 个图像查询对应 5 个文本描述,只要返回 5 个对应文本中的 1 个即可满足要求;而文本检索图像子任务中,1 个文本查询只对应 1 个图像,只有返回对应的那个图像才能满足要求。

近年来,基于分数方法在 MS-COCO 数据集 1K 和 5K 设置下的性能比较见表 3,在 Flickr30K 数据集上的性能比较见表 4。

表 3 近年来基于分数方法在 MS-COCO 数据集上的性能比较

Tab. 3 Performance comparison of score-based methods on the MS-COCO dataset in recent years

年份	方法	MS-COCO 1K							MS-COCO 5K						
		图像检索文本			文本检索图像			R_s	图像检索文本			文本检索图像			R_s
		R_1	R_5	R_{10}	R_1	R_5	R_{10}		R_1	R_5	R_{10}	R_1	R_5	R_{10}	
2017	DVSA[65]	38.4	69.9	80.5	27.4	60.2	74.8	351.2	16.5	39.2	52.0	10.7	29.6	42.2	190.2
2018	SCO[68]	69.9	92.9	97.5	56.7	87.5	94.8	499.3	42.8	72.3	83.0	33.1	62.9	75.5	369.6
2018	SCAN[49]	72.7	94.8	98.4	58.8	88.4	94.8	507.9	50.4	82.2	90.0	38.6	69.3	80.4	410.9
2019	CMR-SC[30]	73.8	95.3	98.3	59.9	88.9	94.9	511.1							
2019	ACMM[69]	84.1	97.8	99.4	60.7	88.7	94.9	525.6	66.9	89.6	94.9	39.5	69.6	81.1	441.6
2019	CAMP[70]	72.3	94.8	98.3	58.5	87.9	95.0	506.8	50.1	82.1	89.7	39.0	68.9	80.2	410.0
2020	ACMNet[75]	72.1	95.2	98.1	59.2	88.1	94.4	507.1							
2020	SGM[71]	73.4	93.8	97.8	57.5	87.3	94.3	504.1	50.0	79.3	87.9	35.3	64.9	76.5	393.9
2020	CAAN[28]	75.5	95.4	98.5	61.3	89.7	95.2	515.6	52.5	83.3	90.9	41.2	70.3	82.9	421.1
2020	IMRAM[81]	76.7	95.6	98.5	61.7	89.1	95.0	516.6	53.7	83.2	91.0	39.7	69.1	79.8	416.5
2020	GSMN[50]	78.4	96.4	98.6	63.3	90.1	95.7	522.5							
2020	Oscar[89]	89.8	98.8	99.7	78.2	95.8	98.3	560.6	73.5	92.2	96.0	57.5	82.8	89.8	491.8
2021	DVSI[80]	75.6	95.2	98.2	58.3	87.0	93.4	507.7							
2021	SGRAF[72]	79.6	96.2	98.5	63.2	90.7	96.1	524.3	57.8		91.6	41.9		81.3	
2021	VinVL[85]	90.8	99.0	99.8	78.8	96.1	98.5	563.0	75.4	92.9	96.2	58.8	83.5	90.3	497.1
2022	CMTFN-SRG[73]	75.6	95.2	98.3	63.0	90.0	95.4	517.5	53.0	81.5	89.7	40.3	71.0	81.7	417.2
2022	L3S-KD[82]	79.8	96.2	98.5	63.5	90.2	95.6	523.7	58.9	84.9	91.7	41.7	71.0	81.3	429.7
2023	HAS[78]	69.6	93.0	97.5	56.3	87.4	94.1	497.9	43.5	75.4	85.5	33.0	64.2	76.5	378.1
2023	CMIIRN[79]	78.3	96.2	98.8	64.2	90.7	96.1	524.3	58.4	84.3	92.0	42.8	72.1	82.7	432.3
2023	BCAN[76]	81.7	98.0	99.2	63.9	91.1	96.4	530.3	60.0	85.7	91.7	40.6	69.4	80.3	427.7
2023	CMSEI[74]	81.4	96.6	98.8	65.8	91.8	96.8	531.1	61.5	86.3	92.7	44.0	73.4	83.4	441.2
2023	CSIC[32]								67.5	89.3	94.3	53.2	80.1	88.1	472.5
2023	PTME[92]								80.1	94.3	97.3	61.8	84.0	90.2	507.7
2024	SAM[91]	81.6	96.7	99.0	68.7	93.6	97.3	536.9	56.8	85.6	92.1	45.2	75.2	85.1	440.0
2024	ESL[77]	84.0	97.2	99.0	67.8	92.7	97.1	537.9	65.8	88.5	94.0	45.7	75.7	85.2	454.9

注:双下划线部分表示 VLP 方法。

表 4　近年来基于分数方法在 Flickr30K 数据集上的性能比较

Tab. 4　Performance comparison of score-based methods on the Flickr30K dataset in recent years

年份	方法	图像检索文本			文本检索图像			R_s
		R_1	R_5	R_{10}	R_1	R_5	R_{10}	
2017	DVSA[65]	22.2	48.2	61.4	15.2	37.7	50.5	235.2
2018	SCO[68]	55.5	82.0	89.3	41.1	70.5	80.1	418.5
2018	SCAN[49]	67.4	90.3	95.8	48.6	77.7	85.2	465.0
2019	CMR-SC[30]	69.7	91.7	96.4	54.0	79.7	87.2	478.7
2019	ACMM[69]	85.2	96.7	98.4	53.8	79.8	86.8	500.7
2019	CAMP[70]	68.1	89.7	95.2	51.5	77.1	85.3	466.9
2020	ACMNet[75]	66.0	90.7	95.8	51.6	78.0	85.8	467.9
2020	SGM[71]	71.8	91.7	95.5	53.5	79.6	86.5	478.6
2020	CAAN[28]	70.1	91.6	97.2	52.8	79.0	87.9	478.6
2020	IMRAM[81]	74.1	93.0	96.6	53.9	79.4	87.2	484.2
2020	GSMN[50]	76.4	94.3	97.3	57.4	82.3	89.0	496.8
2020	Oscar[89]							
2021	DVSI[80]	67.0	90.0	95.3	49.3	76.8	84.3	462.7
2021	SGRAF[72]	77.8	94.1	97.4	58.5	83.0	88.8	499.6
2021	VinVL[85]							
2022	CMTFN-SRG[73]	73.6	91.7	96.3	56.2	82.5	89.4	489.7
2022	L3S-KD[82]	77.3	93.8	97.6	57.6	83.1	89.4	498.8
2023	HAS[78]	64.8	88.3	92.5	49.1	77.6	86.2	458.5
2023	CMIIRN[79]	79.6	95.4	98.2	62.8	87.1	92.0	515.1
2023	BCAN[76]	81.8	96.2	98.1	56.2	83.1	89.4	504.8
2023	CMSEI[74]	82.3	96.4	98.6	64.1	87.3	92.6	521.3
2023	CSIC[32]	88.5	98.1	99.4	75.3	93.6	96.8	551.7
2023	PTME[92]	96.0	99.8	100.0	83.9	96.3	100.0	576.0
2024	SAM[91]	82.0	95.8	98.4	65.1	88.3	93.6	523.2
2024	ESL[77]	84.9	97.0	98.9	67.0	89.2	93.7	530.2

注：双下划线部分表示 VLP 方法。

由表 3 和表 4 可以看出，近年来基于分数方法在 MS-COCO 数据集和 Flickr30K 数据集上的性能都有了稳步提升，且精度相差不大。在包含 1 000 个图像和 5 000 个文本的同样规模的测试集中，对于非 VLP 方法，图像检索文本任务中R_1的最好结果达到了 85.2，R_5的最好结果达到了 98.0，R_{10}的最好结果达到了 99.4；文本检索图像任务中R_1的最好结果达到了 68.7，R_5的最好结果达到了 93.6，R_{10}的最好结果达到了 97.3；R_s的最好结果达到了 537.9。对于 VLP 方法，图像检索文本任务中R_1的最好结果达到了 96.0，R_5的最好结果达到了 99.8，R_{10}的最好结果达到了 100.0；文本检索图像任务中R_1的最好结果达到了 83.9，R_5的最好结果达到了 96.3，R_{10}的最好结果达到了 100.0；R_s的最好结果达到了 576.0，体现出了基于分数方法在跨模态检索任务方面的鲁棒性和稳定性。此外，由于部分方法在 MS-COCO 1K 设置下的测试结果缺失和测试集规模的限制，使得 VLP 方法对比非 VLP 方法的性能优越性（图像检索文本：R_1 96.0 vs. 85.2，R_5 99.8 vs. 98.0，R_{10} 100.0 vs. 99.4；文本检索图像：R_1 83.9 vs. 68.7，R_5 96.3 vs. 93.6，R_{10} 100.0 vs. 97.3；R_s 576.0 vs. 537.9）没有很好地体现出来。在 MS-COCO 5K 设置中，VLP 方法对比非 VLP 方法体现出了极大的优越性（图像检索文本：R_1 80.1 vs. 66.9，R_5 94.3 vs. 89.6，R_{10} 97.3 vs. 94.9；文本检索图像：R_1 61.8 vs. 45.7，R_5 84.0 vs. 75.7，R_{10} 90.3 vs. 85.2；R_s 507.7 vs. 454.9），这是由于 VLP 方法借助 Transformer 在图像和文本之间执行深度交互和预训练的结果。

由表 1 ~ 4 可以看出，自从 2018 年 SCAN 被提出以来，跨模态检索领域涌现了大量以 SCAN 为基

线的基于分数方法，如 CMR-SC、ACMM、CAMP、SGM、GSMN 和 SGRAF 等。在包含 1 000 个图像和 5 000 个文本的同样规模的测试集（Flickr30K 测试集和 MS-COCO 1K 设置）中，由于缺乏特征交互，基于特征方法的性能整体上弱于非 VLP 和 VLP 的基于分数方法的性能（图像检索文本：R_1 84.6 vs. 85.2 vs. 96.0，R_5 96.9 vs. 98.0 vs. 99.8，R_{10} 98.9 vs. 99.4 vs. 100.0；文本检索图像：R_1 66.4 vs. 68.7 vs. 83.9，R_5 92.2 vs. 93.6 vs. 96.3，R_{10} 96.5 vs. 97.3 vs. 100.0；R_s 531.9 vs. 537.9 vs. 576.0），在 MS-COCO 5K 设置下的性能比较结果也类似。因此，近年来，基于分数方法成为跨模态检索任务研究的主流。但是由于基于分数方法存在固有的测试和推理时间复杂度高的问题，构建高精度、高效率的跨模态检索方法逐渐引起了研究人员的关注[92-93]。

4 亟待解决的问题

目前，跨模态检索领域主要存在 3 个关键问题：

1）在单模态特征编码中，不同模态局部特征之间缺乏交互。在现有的基于特征方法中，由于图像与文本之间缺乏交互，使得其无法衡量不同模态局部片段之间的相似性，无法学习图像区域和文本单词之间的对应关系，导致同一时期基于特征方法的检索性能普遍弱于基于分数方法。因此，在不提高测试和推理时间复杂度的前提下，如何在单模态特征编码中利用不同模态局部特征交互增强单模态特征[27,94-96]，是跨模态检索领域亟待解决的关键问题之一。

2）局部特征聚合为全局特征时，存在信息损失。现有的特征聚合方法要么采用局部特征加权组合，没有关注特征向量通道维度的不同权重；要么采用固定池化操作聚合局部特征，没有融合各种测量单元潜在的有判别力的信息，这些都会导致全局特征信息损失，从而导致检索性能次优。因此，如何在聚合全局特征的同时关注向量通道维度的重要成分，为不同特征集合学习最佳池化策略[46,97-98]，是跨模态检索领域亟待解决的关键问题之一。

3）基于分数方法测试和推理阶段时间复杂度高。现有基于分数方法凭借图像和文本局部特征的深度交互，在跨模态检索任务方面取得了优异性能。但是，也正是由于这种深度交互，导致基于分数方法在大规模数据集上的测试和推理时间复杂度非常高，这严重影响了其在真实场景中的应用。因此，在不提高跨模态检索系统测试和推理时间复杂度的前提下，如何合理利用基于分数方法的高精度，如何重新设计基于分数方法的训练和推理策略[49,72,99]，是跨模态检索领域亟待解决的关键问题之一。

5 结 论

图像文本跨模态检索是指给定一种模态的查询，从待检索集中查找与查询相关的另一种模态的样本。本文主要针对面向无标签数据集基于实值特征的图像文本跨模态检索（以下简称跨模态检索）方法展开综述，将现有的跨模态检索方法按照测试和推理时间复杂度分为基于特征方法和基于分数方法分别展开论述。研究的主要结论包括：

1）在包含 1 000 个图像和 5 000 个文本的同样规模的测试集（Flickr30K 和 MS-COCO 1K 设置）中，由于缺乏特征交互，基于特征方法的性能整体上弱于 VLP 和非 VLP 的基于分数方法的性能（图像检索文本：R_1 84.6 vs. 85.2 vs. 96.0，R_5 96.9 vs. 98.0 vs. 99.8，R_{10} 98.9 vs. 99.4 vs. 100.0；文本检索图像：R_1 66.4 vs. 68.7 vs. 83.9，R_5 92.2 vs 93.6 vs. 96.3，R_{10} 96.5 vs. 97.3 vs. 100.0；R_s 531.9 vs. 537.9 vs. 576.0），在 MS-COCO 5K 设置下的性能比较结果也类似。

2）在基于特征方法方面，现有方法的图像和文本模态往往独立编码，然后在特征空间中进行匹配。这种方式忽略了不同模态之间的潜在关联，导致模态间交互不足，进而影响检索精度。此外，现有方法在局部特征聚合时存在信息损失，导致全局特征不够鲁棒，进而影响检索精度。

3）在基于分数方法方面，现有方法凭借图像和文本局部特征的深度交互，在跨模态检索任务方面取得了优异性能。然而，基于分数方法存在测试和推理时间复杂度高的问题，导致其难以在真实场景的大规模数据集上应用。

参考文献

[1] WANG Kaiye, YIN Qiyue, WANG Wei, et al. A comprehensive survey on cross-modal retrieval[EB/OL]. (2016-07-21)[2024-03-17]. https://arxiv.org/pdf/1607.06215

[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. DOI:10.1109/TCSVT.2017.2705068

[3] 彭宇新, 綦金玮, 黄鑫. 多媒体内容理解的研究现状与展望[J]. 计算机研究与发展, 2019, 56(1): 183
PENG Yuxin, QI Jinwei, HUANG Xin. Current research status and prospects on multimedia content understanding[J]. Journal of Computer Research and Development, 2019, 56(1): 183. DOI:10.7544/issn1000-1239.2019.20180770

[4] CHANG Xiaojun, MA Zhigang, LIN Ming, et al. Feature interaction augmented sparse learning for fast kinect motion detection[J]. IEEE

Transactions on Image Processing, 2017, 26(8): 3911. DOI:10.1109/TIP.2017.2708506

[5]JIA Yahui, CHEN Weineng, GU Tianlong, et al. A dynamic logistic dispatching system with set-based particle swarm optimization[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018, 48(9): 1607. DOI:10.1109/TSMC.2017.2682264

[6]SHEN Fumin, SHEN Chunhua, LIU Wei, et al. Supervised discrete hashing[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Boston: IEEE, 2015: 37. DOI:10.1109/cvpr.2015.7298598

[7]ZHANG Lining, WANG Lipo, LIN Weisi. Generalized biased discriminant analysis for content-based image retrieval[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2012, 42(1): 282. DOI:10.1109/TSMCB.2011.2165335

[8]杜锦丰, 王海荣, 梁焕, 等. 基于表示学习的跨模态检索方法研究进展[J]. 广西师范大学学报(自然科学版), 2022, 40(3): 1
DU Jinfeng, WANG Hairong, LIANG Huan, et al. Progress of cross-modal retrieval methods based on representation learning[J]. Journal of Guangxi Normal University(Natural Science Edition), 2022, 40(3): 1. DOI:10.16088/j.issn.1001-6600.2021071302

[9]樊花, 陈华辉. 基于哈希方法的跨模态检索研究进展[J]. 数据通信, 2018(3): 39
FAN Hua, CHEN Huahui. Research on cross-modal retrieval based on hash method[J]. Data Communications, 2018(3): 39. DOI:10.3969/j.issn.1002-5057.2018.03.011

[10]徐文婉, 周小平, 王佳. 跨模态检索技术研究综述[J]. 计算机工程与应用, 2022, 58(23): 12
XU Wenwan, ZHOU Xiaoping, WANG Jia. Overview of cross-modal retrieval technology[J]. Computer Engineering and Applications, 2022, 58(23): 12. DOI:10.3778/j.issn.1002-8331.2205-0160

[11]张博麟, 陈征. 跨模态哈希学习研究进展[J]. 无线通信技术, 2019, 28(4): 35
ZHANG Bolin, CHEN Zheng. A survey on cross-modal hash learning[J]. Wireless Communication Technology, 2019, 28(4): 35. DOI:10.3969/j.issn.1003-8329.2019.04.008

[12]张飞飞, 马泽伟, 周玲, 等. 图文跨模态检索研究进展[J]. 数据采集与处理, 2023, 38(3): 479
ZHANG Feifei, MA Zewei, ZHOU Ling, et al. Recent advances in cross modal image text retrieval[J]. Journal of Data Acquisition and Processing, 2023, 38(3): 479. DOI:10.16337/j.1004-9037.2023.03.001

[13]CHANG E, GOH K, SYCHAY G, et al. CBSA: content-based soft annotation for multimodal image retrieval using bayes point machines[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2003, 13(1): 26. DOI:10.1109/TCSVT.2002.808079

[14]HE Xiaofei, MA Weiying, ZHANG Hongjiang. Learning an image manifold for retrieval[C]//12th Annual ACM International Conference on Multimedia. New York: Association for Computing Machinery, 2004: 17. DOI:10.1145/1027527.1027532

[15]BABENKO A, SLESAREV A, CHIGORIN A, et al. Neural codes for image retrieval[C]//13th European Conference on Computer Vision, ECCV 2014. Zurich: Springer, 2014: 584. DOI:10.1007/978-3-319-10590-1_38

[16]SALTON G. Another look at automatic text-retrieval systems[J]. Communications of the ACM, 1986, 29(7): 648. DOI:10.1145/6138.6149

[17]SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information Processing and Management, 1988, 24(5): 513. DOI:10.1016/0306-4573(88)90021-0

[18]GONZALO J, VERDEJO F, CHUGUR I, et al. Indexing with wordnet synsets can improve text retrieval[C]//Proceedings of the COLING/ACL'98 Workshop on Usage of WordNet for Natural Language Processing Systems. Montreal: [s.n.], 1998: 38

[19]HU Weiming, XIE Nianhua, LI Li, et al. A survey on visual content-based video indexing and retrieval[J]. IEEE Transactions on Systems, Man and Cybernetics Part C (Applications and Reviews), 2011, 41(6): 797. DOI:10.1109/TSMCC.2011.2109710

[20]MCGURK H, MACDONALD J. Hearing lips and seeing voices[J]. Nature, 1976, 264(5588): 746. DOI:10.1038/264746a0

[21]WANG Kaiye, HE Ran, WANG Wei, et al. Learning coupled feature spaces for cross-modal matching[C]//2013 IEEE International Conference on Computer Vision. Sydney: IEEE, 2013: 2088. DOI:10.1109/iccv.2013.261

[22]ZHAI Xiaohua, PENG Yuxin, XIAO Jianguo. Heterogeneous metric learning with joint graph regularization for cross-media retrieval[C]//Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI 2013. Bellevue: AAAI, 2013: 1198

[23]PENG Yuxin, ZHU Wenwu, ZHAO Yao, et al. Cross-media analysis and reasoning: advances and directions[J]. Frontiers of Information Technology and Electronic Engineering, 2017, 18(1): 44. DOI:10.1631/FITEE.1601787

[24]HAUPTMANN A, YAN Rong, LIN Weihao, et al. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news[J]. IEEE Transactions on Multimedia, 2007, 9(5): 958. DOI:10.1109/TMM.2007.900150

[25]GAO Lianli, GUO Zhao, ZHANG Hanwang, et al. Video captioning with attention-based LSTM and semantic consistency[J]. IEEE Transactions on Multimedia, 2017, 19(9): 2045. DOI:10.1109/TMM.2017.2729019

[26]WU Jianlong, LIN Zhouchen, ZHA Hongbin. Joint latent subspace learning and regression for cross-modal retrieval[C]//40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2017. Tokyo: Association for Computing Machinery, 2017: 917. DOI: 10.1145/3077136.3080678

[27]LI Jiangtong, NIU Li, ZHANG Liqing. Action-aware embedding enhancement for image-text retrieval[C]//Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022. Vancouver: AAAI, 2022: 1323. DOI:10.1609/aaai.v36i2.20020

[28]ZHANG Qi, LEI Zhen, ZHANG Zhaoxiang, et al. Context-aware attention network for image-text retrieval[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle: IEEE, 2020: 3533. DOI:10.1109/CVPR42600.2020.00359

[29]QU Leigang, LIU Meng, CAO Da, et al. Context-aware multi-view summarization network for image-text matching[C]//Proceedings of the 28th ACM International Conference on Multimedia. Seattle: Association for Computing Machinery, 2020: 1047. DOI:10.1145/

3394171.3413961

[30] CHEN Hui, DING Guiguang, LIN Zijia, et al. Cross-modal image-text retrieval with semantic consistency [C]//Proceedings of the 27th ACM International Conference on Multimedia. Nice: Association for Computing Machinery, 2019: 1749. DOI:10.1145/3343031.3351055

[31] LI Kunpeng, ZHANG Yulun, LI Kai, et al. Image-text embedding learning via visual and textual semantic reasoning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022: 1. DOI:10.1109/TPAMI.2022.3148470

[32] LIU Zejun, CHEN Fanglin, XU Jun, et al. Image-text retrieval with cross-modal semantic importance consistency [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(5): 2465. DOI:10.1109/TCSVT.2022.3220297

[33] CHEN Shizhe, ZHAO Yida, JIN Qin, et al. Fine-grained video-text retrieval with hierarchical graph reasoning [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 10635. DOI: 10.1109/CVPR42600.2020.01065

[34] FENG Zerun, ZENG Zhimin, GUO Caili, et al. Temporal multimodal graph transformer with global-local alignment for video-text retrieval [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(3): 1438. DOI:10.1109/TCSVT.2022.3207910

[35] JIN Weike, ZHAO Zhou, ZHANG Pengcheng, et al. Hierarchical cross-modal graph consistency learning for video-text retrieval [C]//44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2021. New York: Association for Computing Machinery, 2021: 1114. DOI:10.1145/3404835.3462974

[36] SONG Fuhu, HU Jifeng, WANG Che, et al. Cross-modal audio-text retrieval via sequential feature augmentation [C]//2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML). Shanghai: Association for Computing Machinery, 2023: 298. DOI:10.1145/3590003.3590056

[37] CHAO Yiwen, YANG Dongchao, GU Rongzhi, et al. 3CMLF: three-stage curriculum-based mutual learning framework for audio-text retrieval [C]//Proceedings of 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Chiang Mai: IEEE, 2022: 1602. DOI: 10.23919/APSIPAASC55919.2022.9979989

[38] BAI Ye, YI Jiangyan, TAO Jianhua, et al. Fast end-to-end speech recognition via non-autoregressive models and cross-modal knowledge transferring from BERT[J]. IEEE/ACM Transactions on Audio Speech and Language Processing, 2021, 29: 1897. DOI: 10.1109/TASLP.2021.3082299

[39] 欧卫华, 刘彬, 周永辉, 等. 跨模态检索研究综述[J]. 贵州师范大学学报(自然科学版), 2018, 36(2): 114
OU Weihua, LIU Bin, ZHOU Yonghui, et al. Survey on the cross-modal retrieval research[J]. Journal of Guizhou Normal University (Natural Sciences), 2018, 36(2): 114. DOI:10.16614/j.cnki.issn1004-5570.2018.02.019

[40] RASIWASIA N, COSTA P, COVIELLO E, et al. A new approach to cross-modal multimedia retrieval [C]//Proceedings of the ACM Multimedia 2010 International Conference. Firenze: Association for Computing Machinery, 2010: 251. DOI: 10.1145/1873951.1873987

[41] CHUA T, TANG Jinhui, HONG Richang, et al. NUS-WIDE: a real-world web image database from national university of singapore [C]//ACM International Conference on Image and Video Retrieval. Santorini Island: Association for Computing Machinery, 2009: 368. DOI:10.1145/1646396.1646452

[42] RASHTCHIAN C, YOUNG P, HODOSH M, et al. Collecting image annotations using Amazon′s mechanical turk [C]//Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon′s Mechanical Turk. Los Angeles: Association for Computational Linguistics, 2010: 139

[43] LIN T, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context [C]//13th European Conference on Computer Vision, ECCV 2014. Zurich: Springer, 2014: 740. DOI:10.1007/978-3-319-10602-1_48

[44] YOUNG P, LAI A, HODOSH M, et al. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions [J]. Transactions of the Association for Computational Linguistics, 2014, 2: 67. DOI:10.1162/tacl_a_00166

[45] FAGHRI F, FLEET D, KIROS J, et al. VSE++: improving visual-semantic embeddings with hard negatives [C]//29th British Machine Vision Conference, BMVC 2018. Newcastle: BMVA, 2019

[46] CHEN Jiacheng, HU Hexiang, WU Hao, et al. Learning the best pooling strategy for visual semantic embedding [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 15784. DOI: 10.1109/CVPR46437.2021.01553

[47] ENGILBERGE M, CHEVALLIER L, PEREZ P, et al. Finding beans in burgers: deep semantic-visual embedding with localization [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018: 3984. DOI: 10.1109/CVPR.2018.00419

[48] YAN Shiyang, YU Li, XIE Yuan. Discrete-continuous action space policy gradient-based attention for image-text matching [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 8092. DOI: 10.1109/CVPR46437.2021.00800

[49] LEE K, CHEN Xi, HUA Gang, et al. Stacked cross attention for image-text matching [C]//15th European Conference on Computer Vision, ECCV 2018. Munich: Springer, 2018: 212. DOI: 10.1007/978-3-030-01225-0_13

[50] LIU Chunxiao, MAO Zhendong, ZHANG Tianzhu, et al. Graph structured network for image-text matching [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 10918. DOI:10.1109/CVPR42600.2020.01093

[51] LI Junnan, SELVARAJU R R, GOTMARE A D, et al. Align before fuse: vision and language representation learning with momentum distillation [C]//35th Conference on Neural Information Processing Systems, NeurIPS 2021. San Diego: Neural information processing systems foundation, 2021: 9694

[52] KIROS R, SALAKHUTDINOV R, ZEMEL R. Unifying visual-semantic embeddings with multimodal neural language models [EB/OL]. (2014-11-10) [2024-03-17]. https://doi.org/10.48550/arXiv.

1411.2539

[53] WANG Liwei, LI Yin, LAZEBNIK S. Learning deep structure-preserving image-text embeddings[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 5005. DOI:10.1109/CVPR.2016.541

[54] REN Zhou, JIN Hailin, LIN Zhe, et al. Joint image-text representation by gaussian visual-semantic embedding[C]//24th ACM international conference on Multimedia, MM 2016. Amsterdam: Association for Computing Machinery, 2016: 207. DOI:10.1145/2964284.2967212

[55] WANG Bokun, YANG Yang, XU Xing, et al. Adversarial cross-modal retrieval[C]//25th ACM International Conference on Multimedia, MM 2017. Mountain View: Association for Computing Machinery, 2017: 154. DOI:10.1145/3123266.3123326

[56] GU Jiuxiang, CAI Jianfei, JOTY S, et al. Look, imagine and match: improving textual-visual cross-modal retrieval with generative models[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Salt Lake City: IEEE, 2018: 7181. DOI:10.1109/CVPR.2018.00750

[57] ZHANG Li, WU Xiangqian. Multi-task framework based on feature separation and reconstruction for cross-modal retrieval[J]. Pattern Recognition, 2022, 122:108217. DOI:10.1016/j.patcog.2021.108217

[58] ANDERSON P, HE Xiaodong, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Salt Lake City: IEEE, 2018: 6077. DOI:10.1109/CVPR.2018.00636

[59] KRISHNA R, ZHU Yuke, GROTH O, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123(1): 32. DOI:10.1007/s11263-016-0981-7

[60] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137. DOI:10.1109/TPAMI.2016.2577031

[61] ZHU Hongguang, ZHANG Chunjie, WEI Yunchao, et al. ESA: external space attention aggregation for image-text retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(10): 6131. DOI:10.1109/TCSVT.2023.3253548

[62] 梁彦鹏, 刘雪儿, 马忠贵, 等. 嵌入共识知识的因果图文检索方法[J]. 工程科学学报, 2024, 46(2): 317
LIANG Yanpeng, LIU Xueer, MA Zhonggui, et al. Causal image-text retrieval embedded with consensus knowledge[J]. Chinese Journal of Engineering, 2024, 46(2): 317. DOI:10.13374/j.issn2095-9389.2023.05.28.001

[63] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]//5th International Conference on Learning Representations, ICLR 2017. Toulon: International Conference on Learning Representations, 2017

[64] ZHANG Jin, HE Xiaohai, QING Linbo, et al. Cross-modal multi-relationship aware reasoning for image-text matching[J]. Multimedia Tools and Applications, 2022, 81(9): 12005. DOI:10.1007/s11042-020-10466-8

[65] KARPATHY A, LI Feifei. Deep visual-semantic alignments for generating image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 664. DOI:10.1109/TPAMI.2016.2598339

[66] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 580. DOI:10.1109/CVPR.2014.81

[67] SCHUSTER M, PALIWAL K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673. DOI:10.1109/78.650093

[68] HUANG Yan, WU Qi, SONG Chunfeng, et al. Learning semantic concepts and order for image and sentence matching[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018: 6163. DOI:10.1109/CVPR.2018.00645

[69] HUANG Yan, WANG Liang. ACMM: aligned cross-modal memory for few-shot image and sentence matching[C]//2019 IEEE/CVF International Conference on Computer Vision(ICCV 2019). Seoul: IEEE, 2019: 5773. DOI:10.1109/ICCV.2019.00587

[70] WANG Zihao, LIU Xihui, LI Hongsheng, et al. CAMP: cross-modal adaptive message passing for text-image retrieval[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019). Seoul: IEEE, 2019: 5763. DOI:10.1109/ ICCV.2019.00586

[71] WANG Sijin, WANG Ruiping, YAO Ziwei, et al. Cross-modal scene graph matching for relationship-aware image-text retrieval[C]//2020 IEEE Winter Conference on Applications of Computer Vision(WACV). Snowmass Village: IEEE, 2020: 1497. DOI:10.1109/WACV45572.2020.9093614

[72] DIAO Haiwen, ZHANG Ying, MA Lin, et al. Similarity reasoning and filtration for image-text matching[C]//35th AAAI Conference on Artificial Intelligence. Vancouver: Association for the Advancement of Artificial Intelligence, 2021: 1218

[73] 刘长红, 曾胜, 张斌, 等. 基于语义关系图的跨模态张量融合网络的图像文本检索[J]. 计算机应用, 2022, 42(10): 3018
LIU Changhong, ZENG Sheng, ZHANG Bin, et al. Cross-modal tensor fusion network based on semantic relation graph for image-text retrieval[J]. Journal of Computer Applications, 2022, 42(10): 3018. DOI:10.11772/j.issn.1001-9081.2021091622

[74] GE Xuri, CHEN Fuhai, XU Songpei, et al. Cross-modal semantic enhanced interaction for image-sentence retrieval[C]//23rd IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa: IEEE, 2023: 1022. DOI:10.1109/WACV56688.2023.00108

[75] CHEN Hui, DING Guiguang, LIN Zijia, et al. ACMNet: adaptive confidence matching network for human behavior analysis via cross-modal retrieval[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2020, 16(1): 1. DOI:10.1145/3362065

[76] LIU Yang, LIU Hong, WANG Huaqiu, et al. BCAN: bidirectional correct attention network for cross-modal retrieval[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023: 1. DOI:10.1109/TNNLS.2023.3276796

[77] ZHANG Kun, HU Bo, ZHANG Huatian, et al. Enhanced semantic similarity learning framework for image-text matching[J]. IEEE

Transactions on Circuits and Systems for Video Technology, 2024, 34(4):2973. DOI:10.1109/TCSVT.2023.3307554

[78]杨晓宇, 李超, 陈舜尧, 等. 基于 Transformer 的图文跨模态检索算法[J]. 计算机科学, 2023, 50(4): 141
YANG Xiaoyu, LI Chao, CHEN Shunyao, et al. Text-image cross-modal retrieval based on Transformer[J]. Computer Science, 2023, 50(4): 141. DOI:10.11896/jsjkx.220100083

[79]魏钰琦, 李宁. 用于图文检索的跨模态信息交互推理网络[J]. 计算机工程与应用, 2023, 59(16): 115
WEI Yuqi, LI Ning. Cross-modal information interaction reasoning network for image and text retrieval[J]. Computer Engineering and Applications, 2023, 59(16): 115. DOI:10.3778/j.issn.1002-8331.2205-0056

[80]WU Chunlei, WU Jie, CAO Haiwen, et al. Dual-View semantic inference network for image-text matching[J]. Neurocomputing, 2021, 426: 47. DOI:10.1016/j.neucom.2020.09.079

[81]CHEN Hui, DING Guiguang, LIU Xudong, et al. IMRAM: iterative matching with recurrent attention memory for cross-modal image-text retrieval[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle: IEEE, 2020: 12652. DOI:10.1109/CVPR42600.2020.01267

[82]ZHANG Li, WU Xiangqian. Latent space semantic supervision based on knowledge distillation for cross-modal retrieval[J]. IEEE Transactions on Image Processing, 2022, 31: 7154. DOI:10.1109/TIP.2022.3220051

[83]CAI Guanyu, ZHANG Jun, JIANG Xinyang, et al. Ask & confirm: active detail enriching for cross-modal retrieval with partial query[C]// 18th IEEE/CVF International Conference on Computer Vision, ICCV 2021. Montreal: IEEE, 2021: 1815. DOI:10.1109/ICCV48922.2021.00185

[84]VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//31st Annual Conference on Neural Information Processing Systems, NIPS 2017. Long Beach: Neural Information Processing Systems Foundation, 2017: 5999

[85]ZHANG Pengchuan, LI Xiujun, HU Xiaowei, et al. VinVL: revisiting visual representations in vision-language models[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Nashville: IEEE, 2021: 5575. DOI:10.1109/CVPR46437.2021.00553

[86]ZHOU Luowei, PALANGI H, ZHANG Lei, et al. Unified vision-language pre-training for image captioning and VQA[C]//34th AAAI Conference on Artificial Intelligence, AAAI 2020. New York: AAAI, 2020: 13041

[87]KIM W, SON B, KIM I. ViLT: vision-and-language transformer without convolution or region supervision[C]//38th International Conference on Machine Learning, ICML 2021. Vienna: ML Research Press, 2021: 5583

[88]CHO J, LU Jiasen, SCHWENK D, et al. X-LXMERT: paint, caption and answer questions with multi-modal transformers[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP). Stroudsburg: Association for Computational Linguistics, 2020: 8785

[89]LI Xiujun, YIN Xi, LI Chunyuan, et al. Oscar: object-semantics aligned pre-training for vision-language tasks[C]//16th European Conference on Computer Vision, ECCV 2020. Glasgow: Springer Cham, 2020: 121. DOI:10.1007/978-3-030-58577-8_8

[90]XU Hu, GHOSH G, HUANG Poyao, et al. VLM: task-agnostic video-language model pre-training for video understanding[C]// Findings of the Association for Computational Linguistics, ACL-IJCNLP 2021. Stroudsburg: Association for Computational Linguistics, 2021: 4227

[91]PARK P, JANG S, CHO Y, et al. SAM: cross-modal semantic alignments module for image-text retrieval[J]. Multimedia Tools and Applications, 2024, 83(4): 12363. DOI:10.1007/s11042-023-15798-9

[92]陈曦, 彭姣, 张鹏飞, 等. 基于预训练模型和编码器的图文跨模态检索算法[J]. 北京邮电大学学报, 2023, 46(5): 112
CHEN Xi, PENG Jiao, ZHANG Pengfei, et al. Cross-modal retrieval algorithm for image and text based on pre-trained models and encoders[J]. Journal of Beijing University of Posts and Telecommunications, 2023, 46(5): 112. DOI: 10.13190/j.jbupt.2023-146

[93]BAO Hangbo, WANG Wenhui, DONG Li, et al. VLMO: unified vision-language pre-training with mixture-of-modality-experts[C]// 36th Conference on Neural Information Processing Systems, NeurIPS 2022. New Orleans: Neural Information Processing Systems Foundation, 2022

[94]JI Zhong, LIN Zhigang, WANG Haoran, et al. Multi-modal memory enhancement attention network for image-text matching[J]. IEEE Access, 2020, 8: 38438. DOI:10.1109/ACCESS.2020.2975594

[95]LI Jiangtong, LIU Liu, NIU Li, et al. Memorize, associate and match: embedding enhancement via fine-grained alignment for image-text retrieval[J]. IEEE Transactions on Image Processing, 2021, 30: 9193. DOI:10.1109/TIP.2021.3123553

[96]SONG Ge, WANG Dong, TAN Xiaoyang. Deep memory network for cross-modal retrieval[J]. IEEE Transactions on Multimedia, 2019, 21(5): 1261. DOI:10.1109/TMM.2018.2877122

[97]WEN Keyu, GU Xiaodong, CHENG Qingrong. Learning dual semantic relations with graph attention for image-text matching[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(7): 2866. DOI:10.1109/TCSVT.2020.3030656

[98]WU Yiling, WANG Shuhui, SONG Guoli, et al. Learning fragment self-attention embeddings for image-text matching[C]//27th ACM International Conference on Multimedia(MM). Nice: Association for Computing Machinery, 2019: 2088. DOI:10.1145/3343031.3350940

[99]CHEN Yenchun, LI Linjie, YU Licheng, et al. UNITER: universal image-text representation learning[C]//16th European Conference on Computer Vision. Glasgow: Springer, 2020: 104. DOI:10.1007/978-3-030-58577-8_7