计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20180770
56(1):183-208, 2019

多媒体内容理解的研究现状与展望

彭宇新　綦金玮　黄　鑫
(北京大学计算机科学技术研究所　北京　100871)
(pengyuxin@pku.edu.cn)

Current Research Status and Prospects on Multimedia Content Understanding

Peng Yuxin, Qi Jinwei, and Huang Xin
(*Institute of Computer Science and Technology, Peking University, Beijing* 100871)

Abstract　With the rapid development of multimedia and Internet technologies, a large amount of multimedia data has been rapidly emerging, such as image, video, text and audio. Data of different media types from multi-source is heterogeneous in the form but relevant in the semantic. As indicated in the research of cognitive science, the perception and cognition of the environment is through the fusion across different sensory organs of human, which is decided by the human brain's organization structure. Therefore, it has been a key challenge to perform data semantic analysis and correlation modeling across different media types, for achieving comprehensive multimedia content understanding, which has drawn wide interests of both academic and industrial areas. In this paper, the basic concepts, representative methods and research status of 5 latest highlighting research topics of multimedia content understanding are referred, including fine-grained image classification and retrieval, video classification and object detection, cross-media retrieval, visual description and generation, and visual question answering. This paper further presents the major challenges of multimedia content understanding, as well as gives the development trend in the future. The goal of this paper is to help readers get a comprehensive understanding on the research status of multimedia content understanding, draw more attention of researchers to relevant research topics, and provide the technical insights to promote further development of this area.

Key words　multimedia content understanding; fine-grained image classification and retrieval; video classification and object detection; cross-media retrieval; visual description and generation; visual question answering

摘　要　随着多媒体和网络技术的迅猛发展,海量的图像、视频、文本、音频等多媒体数据快速涌现.这些不同媒体的数据在形式上多源异构,语义上相互关联.认知科学研究表明,人脑生理组织结构决定了其对外界的感知和认知过程是跨越多种感官信息的融合处理.如何对不同媒体的数据进行语义分析和关联建模以实现多媒体内容理解,成为了一个研究和应用的关键问题,受到了学术界和工业界的广泛关注.选取了多媒体内容理解的5个最新热点研究方向:图像细分类与检索、视频分类与目标检测、跨媒体检索、视觉描述与生成、视觉问答,分别阐述了它们的基本概念、代表性方法、研究现状等,并进一步阐述了多媒体内容理解面临的重要挑战,同时给出未来的发展趋势,旨在帮助读者全面了解多媒体内容理解的研究现状,吸引更多研究人员投身相关研究并为他们提供技术参考,推动该领域的进一步发展.

收稿日期:2018-11-16;**修回日期**:2018-12-04
基金项目:国家自然科学基金项目(61771025,61532005)
This work was supported by the National Natural Science Foundation of China (61771025, 61532005).
通信作者:彭宇新(pengyuxin@pku.edu.cn)

关键词 多媒体内容理解;图像细分类与检索;视频分类与目标检测;跨媒体检索;视觉描述与生成;视觉问答

中图法分类号 TP391

随着多媒体和网络技术的迅猛发展,海量的图像、视频、文本、音频等多媒体数据快速涌现.根据Gartner统计,图像、视频数据已经占到大数据的90%以上.2017年全球图像、视频等多媒体数据总量已经超过20 ZB.这些数据在形式上多源异构,语义上相互关联,对于感知与认知客观世界至关重要.认知科学研究表明:人脑生理组织结构决定了其对外界的感知和认知过程是跨越多种感官信息的融合处理[1].而如何对这些多媒体数据进行语义分析和关联建模以实现多媒体内容理解,就成为了研究与应用的关键问题[2].

多媒体内容理解在互联网内容监测、态势分析、智能医疗、智慧城市等重要领域具有广阔的应用前景.2017年7月国务院印发的《新一代人工智能发展规划》将图像、视频分析技术列为社会综合治理、新型犯罪侦查、反恐等迫切需求的关键技术,彰显了多媒体内容理解对于保障国家安全和社会稳定的重要意义.同时,多媒体内容理解也受到了Google、Facebook、Microsoft、IBM、百度、阿里、腾讯等著名企业的广泛关注,投入了大量资源进行相关研究与应用工作.这表明多媒体内容理解既符合国家的战略需求,也切合企业的市场需求,正深刻地改变着我们的社会与生活方式.

多媒体数据具有语义抽象、类别细化、非结构化、数据量大的特点,使得多媒体内容理解面临着两大难题:"异构鸿沟"和"语义鸿沟"."异构鸿沟"是指图像、视频等不同媒体数据的表征不一致,导致难以实现多种媒体数据的统一表征与综合利用."语义鸿沟"是指对于多媒体数据中的每一种媒体如图像,存在数据表征和人类认知之间的矛盾.传统的单一媒体分析方法因为信息有限而难以实现内容理解,因此如何综合利用多媒体数据缩短"异构鸿沟"和"语义鸿沟",成为了多媒体内容理解研究的关键挑战.多媒体内容理解吸引了国际学术界的广泛关注,每年在TPAMI,IJCV,TIP,TMM,TCSVT,TOMM,NIPS,ICML,CVPR,ICCV,ACM MM等国际知名期刊和会议上都有大量论文发表.多媒体内容理解是一个比较广泛的研究方向,不仅需要对图像、文本等单一媒体数据进行分析,也需要对多种媒体数据进行综合分析以实现语义协同,涉及多媒体、计算机视觉、人工智能、机器学习、模式识别等多个领域的知识.近年来,深度学习的兴起和发展为多媒体内容理解提供了新的方法和模型,在研究与应用上都取得了显著进展.代表性研究工作包括图像细分类与检索、视频分类与目标检测、跨媒体检索、视觉描述与生成、视觉问答等,简单叙述如下:

1) 在图像细分类与检索上,面对类别细化的海量图像数据,如何辨识图像细粒度差异、建立高效的索引结构,实现图像数据的准确理解与快速检索,是大数据有效利用急需解决的关键问题,具有重要的研究和应用价值.此外,如何引入文本信息综合建模图像和文本数据,促进图像语义的精细理解,是图像细分类方向研究的前沿问题.

2) 在视频分类与目标检测上,如何有效建模视频的时空结构和运动信息,如何处理视频中运动模糊、视角变化、遮挡、形变等复杂情况,建立高效率、可扩展的视频分类与目标检测模型,是视频分析方向研究的重要问题.

3) 在跨媒体检索上,如何借鉴人脑感知与认知的跨媒体特性,通过跨媒体数据的统一表征等方法,突破"异构鸿沟"导致的相似性难以度量的问题,满足用户对不同媒体数据交叉检索的需求,是实现跨媒体数据语义互通与理解的关键问题.

4) 在视觉描述与生成上,如何实现图像视频与自然语言之间的相互转化与生成,一方面使计算机自动生成视觉内容的自然语言描述,另一方面让计算机能根据人类自然语言描述自动生成图像和视频,成为了跨越计算机视觉与自然语言理解两大研究领域的前沿问题.

5) 在视觉问答上,如何正确理解文本问题的意图和视觉内容中的对象种类、关系等信息,对于实现计算机准确回答问题至关重要.同时,由于问题具备自由开放的特性,如何引入常识和背景知识进行推理,也成为了提高视觉问答系统智能化程度的一个关键挑战.作为一种高层的跨媒体语义理解问题,视觉问答是近年来一个新的研究热点和难点,有望形成新型的人机交互方式,具有广泛的潜在应用价值.

虽然近年来多媒体内容理解取得了一系列进展,但仍然面临重要挑战.例如跨媒体推理、小样本训练与学习、无监督条件下的多媒体内容理解、跨媒体

知识图谱、跨媒体数据相互生成、视觉知识嵌入与推理、多媒体内容理解的实际应用等.本文在第6节中给出了一些思考和看法.

本文从图像细分类与检索、视频分类与目标检测、跨媒体检索、视觉描述与生成、视觉问答这5个研究方向出发,分别阐述其基本概念与代表性方法.此外,进一步阐述了多媒体内容理解所面临的重要挑战,并给出发展趋势.本文旨在帮助读者了解多媒体内容理解的研究现状,吸引更多研究人员投身相关研究并为他们提供技术参考①,推动该领域的进一步发展.

1 图像细分类与检索

随着图像数据的快速增长,图像细分类与检索成为了用户迫切需要的关键技术,也是计算机视觉、模式识别领域的重要研究方向,对于实现多媒体大数据的有效利用具有重要意义.

1.1 图像细分类

图像分类旨在使计算机自动识别图像内容的语义类别.传统的图像分类一般仅识别图像中是否包含某个大类对象(如狗、鸟等),而图像细分类聚焦于大类中细粒度子类的精细识别(如识别狗这一大类中的阿拉斯加犬、哈士奇等子类,又如识别鸟这一大类中的上百个子类),在医疗检测、无人驾驶、动植物保护、海洋作业等领域具有重要的应用价值.

与传统的图像分类相比,图像细分类具有更大的挑战,主要体现在类间差异小和类内差异大.如图1所示都是鸟类,左边两幅图像分别属于"great crested flycatcher"和"acadian flycatcher"两个不同的子类,但颜色、外观等却很相似;而右边两幅图像同属于子类"indigo bunting",但由于姿态、光照等不同导致外观差异很大.不同细粒度子类的差异主要位于对象的部件中,例如鸟的头部、躯干、脚部等.因此,如何检测并学习对象及其关键部件成为了图像细分类的关键问题.

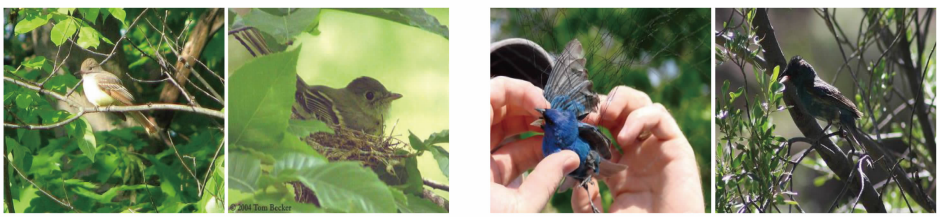

Fig. 1 Examples of fine-grained images

图1 细粒度图像示例

传统方法主要依赖于手工设计的特征,但由于手工特征的表达能力有限,分类效果并不理想.随着深度学习的发展与应用,图像细分类也取得了很大突破.根据使用的标注信息不同,本节将从3个方面对图像细分类进行介绍:强监督图像细分类、弱监督图像细分类、基于属性和文本描述的图像细分类.

1.1.1 强监督图像细分类

强监督图像细分类是指除了使用图像级类别标注信息以外,还使用了对象区域(bounding box)、部件位置(part location)等人工标注信息.其中,对象区域和部件位置的标注尤其耗时耗力,标注成本巨大,限制了这类方法的实际应用.

图像细分类的关键在于对象及其部件区域的检测和特征表示.大部分强监督图像细分类方法的思路是利用对象区域和部件位置的标注信息训练相应的检测器,以获得图像中的对象和部件区域并进行特征提取,最后训练分类器进行图像细分类任务.其中,一个经典的方法是Zhang等人提出的part-based R-CNN[3].方法如图2所示,首先使用对象区域和部件位置的标注信息训练R-CNN[4]检测器,利用检测器对使用选择搜索算法(selective search)[5]生成的图像块进行筛选,最后提取对象、部件所对应的图像块特征并训练分类器.

Huang等人提出了一种部件堆叠的卷积神经网络(part-stacked CNN)[6],在定位对象和部件区域之后,提取对象和部件的卷积特征并融合进行分类.该方法利用了对象区域的标注信息对输入图像进行裁剪.

① 本文部分相关工作的代码请访问 https://github.com/PKU-ICST-MIPL

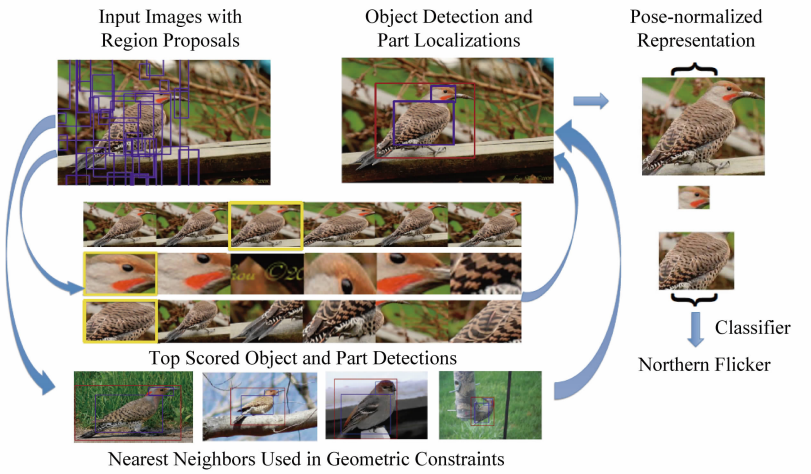

Fig. 2 Method of part-based R-CNN[3]

图 2 part-based R-CNN 方法[3]

除了上述方法之外,还有很多强监督图像细分类方法[7-8].借助于人工标注信息,这类方法取得了较高的细分类准确率.但是,因为标注成本巨大导致这类方法很难进行实际应用.如何在仅使用图像类别标注的条件下进行图像细分类,是图像细分类走向实际应用的关键问题.

1.1.2 弱监督图像细分类

弱监督图像细分类是指仅使用图像类别标注信息,不依赖于人工标注的对象区域和部件位置信息.其挑战在于如何自动获得图像中对象和部件等关键区域信息.Xiao 等人提出了"对象-部件"两级注意力模型[9],这是首个在训练和测试 2 个阶段均不使用对象区域、部件位置等人工标注信息的图像细分类方法,并且取得了很好的细分类效果[10].该工作提出的"弱监督深度图像细分类"已经发展成为该领域的一个新研究方向,受到该工作启发,其他弱监督图像细分类方法陆续被提出[11-13].."对象-部件"两级注意力模型主要包含 2 个部分:1)对象级注意力模型.用于筛选包含对象区域的图像块并进行分类.首先,通过选择搜索算法产生候选图像块,然后利用训练的卷积神经网络自动筛选出与对象类别相关的图像块,并提取深度特征训练对象级分类器.2)部件级注意力模型.用于检测包含对象不同部件的图像块并进行分类.首先,对于对象级注意力模型中特定卷积层的卷积核生成相似度矩阵,并对相似度矩阵进行谱聚类以生成具有不同部件语义的卷积核组(即部件检测器).然后,利用部件检测器对选择搜索算法生成的候选图像块进行预测,选择出其中包含对象部件的图像块,并提取深度特征训练部件级分类器.最后,将对象级分类器得分和部件级分类器得分进行加权融合得到最终的图像细分类得分.在此基础上,Peng 等人进一步提出了空间拓扑注意力学习方法[14].在对象级注意力模型上,通过对输入图像的显著性分布进行自动统计,自动定位图像中的对象区域并去除背景信息,使得模型聚焦于对象特征的学习.在部件级注意力模型上,利用视觉对象区域与显著性部件之间以及显著性部件相互之间的空间关系约束,实现了显著性部件的有效选择.这样通过两级注意力驱动的联合学习,提高了图像细分类的准确率.此外,Zhang 等人[12-13]利用卷积网络中卷积核的选择特性,选取对对象部件具有强响应的卷积核作为部件检测器,并通过正则化多示例学习训练检测器以提高定位能力.Fu 等人提出递归注意力卷积神经网络(recurrent attention convolutional neural network)[15],利用多尺度子网络对显著性区域注意力和区域特征表示进行递归学习,最终将多个尺度的子网络卷积特征融合实现图像细分类任务.Wang 等人[16]在卷积神经网络中学习辨识性过滤器来提取图像中与类别相关的区域,以提高图像细分类的准确率.

上述方法对于显著性部件的定位以及数目的设定依赖于先验知识和实验验证,为了能够自适应地学习显著性部件的定位和数目,He 等人提出了堆叠式深度强化学习方法[17],通过对象-部件两阶段的强化学习,序列式地定位对象及其显著性部件,在此过程中显著性部件的数目由注意力奖惩和语义奖惩函数的反馈自动学习得到.为了加速图像细分类方法的速度,He 等人[18-19]在只使用图像级别的类别标注的情况下,提出了显著性引导的快速定位方法,能够在取得更高细分类准确率的同时,加快定位

与细分类速度.首先,通过多级注意力驱动的定位学习自动定位图像中不同的显著性区域,这些区域提供互补且互不冗余的信息以提升细分类准确率.其次,通过多路显著性定位网络,使用权值共享能够在一次前向传播中同时生成多个显著性区域,避免了多个区域生成的重复性计算,从而加快了速度.除了上述介绍的方法以外,还有很多弱监督图像细分类方法,如 Zhang 等人[20]直接从卷积特征中挑选辨识性部件特征以提高细分类准确率,Lin 等人[21]设计的端到端双线性 CNN 和 Zhao 等人[22]提出的代价敏感的深度度量学习方法等.

1.1.3 基于属性和文本描述的图像细分类

为了更好地理解图像内容,研究者开始利用图像的属性标注(如"锥形的鸟嘴"、"蓝色的翅膀"等)和文本描述信息(如"这是一只有红色鸟嘴和白色翅膀的鸟,其在海面上飞翔")来学习图像的特征表示.Zhou 等人[8]利用属性标注信息来构建二分图,通过引入二分图标签训练卷积神经网络.Liu 等人[23]提出了一种属性驱动的注意力定位方法,通过在强化学习中引入属性信息来定位显著性区域.Chen 等人[24]利用属性标注构建知识图谱(knowledge graph),Xu 等人[25]引入外部知识库构建知识图谱,进一步优化了图像的特征表示学习.

此外,综合建模多种媒体数据并学习媒体间的关联知识,能够进一步增强单媒体的表示能力.He 等人[26]引入了文本描述信息,提出了视觉-文本多源语义嵌入的视觉表示方法,采用卷积-循环网络(convolutional recurrent net, CNN-RNN)对图像和文本进行联合建模,构建相容函数(compatibility function)以联合考虑视觉、文本提供的多源语义信息,利用二者的差异性、互补性学习更具辨识力的视觉表示.通过多源、多粒度、多层次的视觉对象描述,进一步提高了图像细分类的准确率.针对只有一个训练样本条件下的图像细分类问题,He 等人利用文本信息,提出了基于选择与生成的数据增广方法[27],通过多示例学习与生成式对抗学习,对数据进行分割、过滤、再选择和生成,利用生成式对抗网络扩充图像训练数据的多样性,实现了单训练样本条件下的图像细分类.

近期,研究者开始进行视频细分类的研究工作.例如 Zhu 等人[28]构建了 2 个用于视频细分类的数据集,并提出了一种冗余减少的注意力机制用于视频细分类.目前这个工作刚刚开始,有望吸引更多的研究人员加入,类似从图像分类到图像细分类的大量研究,更加符合人类对图像内容细粒度理解的需要.

1.2 图像检索

图像检索是指将一张图像作为查询,检索出图像库中与查询图像相似的图像,并按相似度从大到小排序,在互联网搜索引擎、新闻出版等行业有着广泛的应用.为了满足快速检索的需求,研究者们提出了基于 Hash 的图像检索方法,通过 Hash 函数将图像的高维特征映射到汉明空间,利用二进制汉明编码实现图像的快速检索,同时能节省大量的存储空间.

根据是否使用标注信息,图像 Hash 方法可以分为有监督和无监督 2 种.此外,随着深度学习在图像识别上的突破性进展,深度图像 Hash 方法也被提出并取得了较大进展.下面分别对无监督图像 Hash 方法、有监督图像 Hash 方法和深度图像 Hash 方法进行介绍.

1.2.1 无监督图像 Hash 方法

无监督图像 Hash 方法不依赖训练图像的标注信息(如图像类别标签)来实现 Hash 码的生成.根据是否需要分析图像数据的性质(如数据流形结构),无监督图像 Hash 方法可以进一步分为数据独立的无监督图像 Hash 方法和数据依赖的无监督图像 Hash 方法.数据独立的无监督图像 Hash 方法不针对任何数据和任务设计 Hash 函数;数据依赖的无监督图像 Hash 方法需要根据所使用数据的性质设计相应的 Hash 函数.由于结合了数据的性质,数据依赖的无监督图像 Hash 方法通常比数据独立的无监督图像 Hash 方法在检索任务上有着更高的准确率.

Gionis 等人提出的局部敏感 Hash(locality sensitive hashing, LSH)[29]是经典的数据独立的无监督图像 Hash 方法.LSH 方法将原始空间中相近的数据,以较大的概率映射到同一个 Hash 桶中,将原始空间中高维特征的检索问题转化为汉明空间中 Hash 编码的检索问题.Lü 等人提出 multi-probe LSH[30]方法,通过多次探测的机制降低了 LSH 方法对空间的需求.此外,Satuluri 等人提出 Bayesian LSH[31]方法,结合了贝叶斯推断和 LSH 方法,同时提高了检索任务的准确率和召回率.

数据依赖的无监督图像 Hash 方法通过分析图像数据的性质(如数据流形结构),学习得到更加有效的 Hash 函数.例如 Liu 等人提出 AGH(anchor graph hashing)[32],通过图模型建模数据的近邻结构学习 Hash 函数.此外,还有许多相关的数据依赖的无监督图像 Hash 方法,从不同方面分析数据的

性质,学习更加有效的 Hash 函数. 例如,SH(spectral hashing)[33],MH(manifold hashing)[34],ITQ(iterative quantization)[35]等.

1.2.2 有监督图像 Hash 方法

有监督图像 Hash 方法利用额外的数据标注信息辅助 Hash 函数的学习. 根据标注信息的形式,可以分为基于点标注信息的图像 Hash 方法(pointwise)、基于成对标注信息的图像 Hash 方法(pairwise)、基于三元组标注信息的图像 Hash 方法(tripletwise)和基于排序标注信息的图像 Hash 方法(listwise). 基于点标注信息的图像 Hash 方法[36-37]利用单个数据的标注学习 Hash 函数,但这些标注信息不能表达数据之间的关系,使得方法的检索准确率有限. 因此,基于成对标注信息的图像 Hash 方法被提出,其利用了数据之间的关系(如相似性、距离等)来实现 Hash 函数的学习. Liu 等人提出的 KSH(kernel-based supervised hashing)[38]方法利用了成对数据的相似性信息,通过将传统的汉明距离计算替换为等价的向量内积计算,以避开对非连续的汉明距离的优化,并提出一种基于贪心的 Hash 函数学习方法. 基于成对标注信息的图像 Hash 方法还有 LAMP(label regularized max-margin partition)[39],ML(metric learning based hashing)[40],MLH(minimal loss hashing)[41]等. 基于三元组标注信息的图像 Hash 方法通常将数据组织成为三元组的形式:

$$\varepsilon=\{x_i,x_i^+,x_i^-\mid sim(x_i,x_i^+)>sim(x_i,x_i^-)\},\quad(1)$$

其中,x_i,x_i^+ 是相似样本对,而 x_i,x_i^- 是不相似样本对,sim 表示两者的相似度. 利用上述三元组标注信息,Li 等人提出的 CGHash(column generation hashing)[42]方法学习得到带权汉明距离函数,使得生成的 Hash 码保持上述三元组标注信息的性质. 此外,相关的工作还有 hamming metric learning[43],OPH(order preserving hashing)[44]等. 基于排序标注信息的图像 Hash 方法,可以看作是基于三元组标注信息的图像 Hash 方法的扩展. 如 Wang 等人提出的 RSH(ranking-based supervised hashing)[45]方法,将排序标注信息分解为一个基于三元组的张量矩阵,类似基于三元组标注的图像 Hash 方法的思想,使生成的 Hash 码保持排序标注信息的性质.

1.2.3 深度图像 Hash 方法

深度图像 Hash 方法借助深度模型在图像特征学习上的优势,实现更有效的 Hash 函数学习[46]. 早期的深度图像 Hash 方法使用了深度生成模型实现 Hash 函数的学习,并取得了一定的效果. 例如多层 RBMs(restricted Boltzmann machines)[47]方法,其框架包括 2 个阶段:1)利用无监督方法,使用训练数据初始化网络参数;2)使用标注信息对得到的网络参数进行调优. 此后,卷积神经网络(convolutional neural network, CNN)在图像分类领域取得了突破性进展,使得一些研究者开始利用 CNN 网络实现更好的 Hash 函数学习. Xia 等人提出的 CNNH(convolutional neural network hashing)[48]方法将卷积神经网络引入到 Hash 函数学习. 该方法分为 2 个阶段:1)利用传统的图像 Hash 方法(如上文提到的 KSH 方法)为训练集生成 Hash 码;2)使用第 1 阶段生成的 Hash 码作为标签,训练得到新的卷积神经网络. CNNH 两阶段的学习分离了 Hash 函数的学习和深度特征的学习,这一定程度上影响了模型在检索任务上的效果. 此后,Lai 等人提出 DNNH(deep neural network hashing)[49]方法,将 Hash 函数和深度特征的学习集成在同一个框架中,并使用三元组标注信息进行模型学习. Li 等人提出了 DPSH(deep pairwise-supervised hashing)[50],使用成对标注信息实现了 Hash 函数和深度特征的同时学习. 此外,Liu 等人提出了 DSH(deep supervised hashing)[51],通过添加正则化项,缓解了因激活函数饱和导致梯度过小带来的训练困难问题,保证了训练的稳定性. Cao 等人提出了 HashNet[52],对于 Hash 方法因量化引起的非凸优化难题,利用平滑的 tanh 函数拟合非连续的符号量化函数 sgn,取得了深度 Hash 优化方法上的进展. 此外,Hash 码的离散性导致在根据汉明距离检索时,返回结果中存在大量样本与查询图像具有相同的汉明距离,无法进一步区分相互的排序关系,导致检索结果不够准确. 针对此问题,Zhang 等人提出的查询自适应的深度权重图像 Hash 方法 QaDWH(query-adaptive deep weighted hashing)[53],在训练阶段设计了 Hash 权重层,计算得到的 Hash 权重向量用于区分不同比特位的 Hash 码对检索结果的重要程度;在查询阶段,提出了查询自适应的图像检索方法,结合图像的类别概率向量和上述 Hash 权重向量,根据查询图像自适应地计算带权汉明距离,增强查询结果相互之间的区分度,提高 Hash 检索的准确率.

深度网络的训练往往需要大量的训练样本,而样本的标注耗时耗力. 因此,一些工作开始关注如何降低对标注信息的依赖. Zhang 等人提出了 SSDH(semi-supervised deep hashing)[54]方法将半监督学习引入到深度图像 Hash 方法中:对于有标注的数据,

采用三元组损失函数用于网络训练;对于无标注的数据,设计了半监督嵌入损失和伪标签损失用于网络训练.半监督嵌入损失针对数据中的流形结构进行建模,为无标注数据的训练提供了有效的监督信息;而伪标签损失利用预测得到的标签,用于最大化模型后验估计以增强模型泛化性.该方法充分利用有标注和无标注的数据同时进行模型训练,能够实现更加有效的 Hash 函数学习,提高检索任务的准确率.

2 视频分类与目标检测

如何准确、高效地分析和理解视频内容并获取有用信息,对于满足用户的信息获取需求至关重要.视频分类和目标检测是视频内容理解的重要问题,下面分别进行阐述.

2.1 视频分类

视频分类旨在使计算机自动识别视频类别,可广泛应用于视频监控、人机交互等.视频具有数据量大、时空信息复杂的特点,给视频分类带来了很大挑战.

在深度学习兴起之前,传统的视频分类方法致力于设计手工特征对视频进行表示.例如,3D Harris[55]检测子将视频抽象为 2D 空间信息和一维时序信息的组合,用时空兴趣点来描述视频 3D 结构中变化显著的局部特征.研究者也对视频中的静态信息和运动信息分别进行建模.例如,方向梯度直方图(histograms of oriented gradient, HOG)[56]计算并统计视频帧局部区域梯度方向的直方图以描述视频静态信息,光流直方图(histograms of optical flow, HOF)[57]计算并统计光流方向的直方图来描述视频中的运动信息,运动边界直方图(motion boundary histograms, MBH)[58]计算并统计水平及垂直方向光流分量灰度图的方向梯度直方图以描述视频的运动信息.

为了利用 HOG,HOF 和 MBH 特征之间的互补性,Wang 等人先后提出了稠密轨迹(dense trajectories, DT)[59]及其改进方法 IDT(improved dense trajectories)[60],通过融合上述多种特征来增强视频特征的描述能力.具体而言,DT 方法通过特征点稠密采样、轨迹跟踪、基于轨迹的特征提取和特征编码 4 个步骤,将 HOG,HOF 和 MBH 等特征有效地融合起来.IDT 方法在 DT 方法基础上进行了改进,通过估计相机运动消除背景光流,并采用费雪向量(Fisher vector, FV)代替词袋编码模型,进一步提升了视频特征的描述能力.

随着深度学习的兴起,基于深度特征的视频分类方法也取得了显著的进展,包括双流卷积神经网络[61]、3D 卷积神经网络[62]等在内的众多模型被提出,逐渐成为了目前的主流方法.双流卷积神经网络[61]由 Simonyan 等人于 2014 年提出,以分离的方式对视频中的静态和运动信息进行建模.如图 3 所示,该模型包括空域 CNN 和时域 CNN 两个分支,分别以视频帧和光流作为输入来建模视频中的静态和运动信息.由于时域 CNN 以光流作为输入,无法在大规模图像数据集上进行预训练,容易导致过拟合现象.为此,Simonyan 等人引入多任务学习策略,使用多个视频数据集训练时域 CNN,有效提高了模型的分类准确率.双流卷积神经网络在视频分类的标准数据集 UCF101 和 HMDB51 上首次取得了优于手工特征的分类准确率,启发了一系列后续工作.Wang 等人[63]针对双流卷积神经网络仅处理单视频帧和短时堆叠光流而无法捕获长时序信息的问题,提出了时序分割网络(temporal segmentation network, TSN).TSN 模型将输入视频划分为若干片段,对每个片段进行稀疏采样和分类,并将分类得分融合得到最终的视频分类结果.由于视频片段的时间跨度较大,TSN 可以对长时序信息进行建模.同时,TSN 通过对视频进行稀疏采样,提高了运算效率.

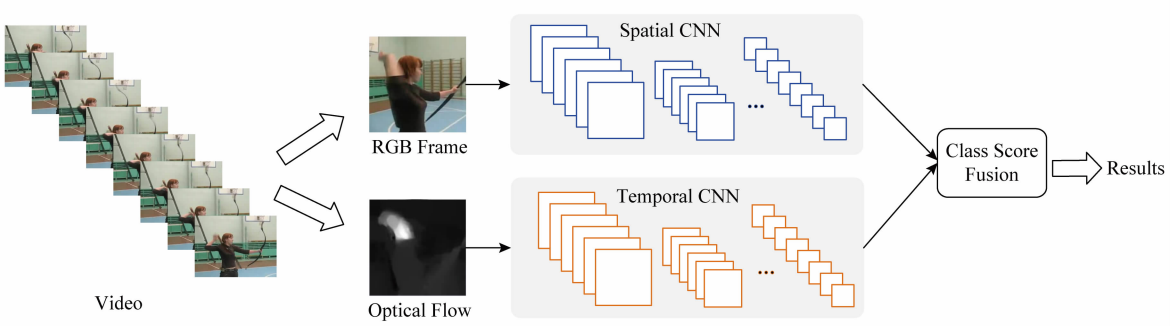

Fig. 3 Two-stream convolutional networks

图 3 双流卷积神经网络模型示意图

Peng 等人[64]将双流卷积神经网络和时空注意力机制相结合,提出了时空注意力双流协同学习(two-stream collaborative learning with spatial-temporal attention, TCLSTA)模型. TCLSTA 利用 CNN 和 LSTM(long shot-term memory)对时空注意力进行联合建模,通过视频帧显著区域定位以及关键帧选择,学习显著的静态和运动特征. 然后进一步对静态、动态特征进行协同优化和自适应学习,提升了视频分类准确率.

3D 卷积神经网络(3D CNN)将图像处理领域的 2D 卷积与池化操作扩展到 3D,利用 3D 卷积核对连续视频帧进行时间维度和空间维度的卷积操作,以实现对视频时空特征的建模. C3D 模型[62]是 3D 卷积神经网络的一个代表性模型,由若干堆叠的 3D 卷积层构成,卷积核大小为 3×3×3,在多个数据集上取得了良好的视频分类准确率. 然而,C3D 以稠密堆叠的视频帧作为输入,且依赖大规模有标注视频数据进行预训练,受计算资源和标注数据的限制较大.

Qiu 等人[65]提出了伪 3D 残差网络(pseudo-3D residual networks),将 3×3×3 的 3D 卷积分解为一个 1×3×3 的 2D 空域卷积和一个 3×1×1 的一维时域卷积,然后将两者组合起来构成伪 3D 卷积结构. 这保证了伪 3D 残差网络相比于同等深度的 2D 卷积神经网络只增加了一定数量的一维卷积结构,避免了参数量的过度增长. 同时,伪 3D 残差网络中的 2D 卷积核可以利用图像数据进行预训练,缓解了网络对于大量有标注视频数据的依赖.

双流卷积神经网络通过光流建模视频的运动信息,十分消耗计算及存储资源. 3D 卷积神经网络在空间和时间维度上同时进行卷积操作,忽略了视频空间和时序信息之间的差异性. 为此,一些研究者探索了其他运动信息建模的方法. 例如,Zhao 等人[66]将视频的卷积特征解耦成静态、运动和表观变化 3 个分支,从 RGB 帧中推导类似光流的运动特征,避免了光流计算带来的额外开销. Sun 等人[67]从光流的定义出发,提出了一种光流引导特征(optical flow guided feature, OFF). 该特征由帧特征图的空间梯度和时间梯度组成,可以直接从 RGB 帧中获取,保证了计算效率,并提升了运动信息建模的准确性.

受到循环神经网络(recurrent neural network, RNN)在自然语言处理、语音识别等领域成功应用的启发,研究者也尝试将其应用于视频时序建模上. 此类方法通常首先用 CNN 提取视频帧特征,然后将这些特征按照时间顺序输入到 RNN 中进行时序建模. LSTM 则是常用的建模视频长时依赖的 RNN 模型. 例如,Ng 等人[68]利用 LSTM 对视频帧和光流特征进行融合,通过实验验证了 LSTM 网络对于光流噪声的鲁棒性以及用 LSTM 实现视频序列特征融合的有效性.

除了上述深度视频分类模型之外,研究者也将不同的网络模型进行结合,或者设计面向视频分类的特定网络模块,以获得更大的性能提升. Carreira 等人[69]将双流卷积神经网络结构和 3D 卷积结合起来,提出了双流 I3D 网络(two-stream inflated 3D ConvNet). 该模型基于 GoogLeNet 中的 Inception-v1 模块,将 2D 卷积层通过参数复制的方式扩展到 3D. 这种扩展方式的一个显著优点是可以利用预训练的 2D 卷积神经网络参数进行初始化,有效缩短了模型的训练时间. Wang 等人[70]设计了面向视频分类的 SMART(simultaneously model appearance and relation)模块,并以此构建了 ARTNet (appearance-relation network)模型,旨在从视频帧序列中直接学习时空特征,而不需要光流信息. SMART 模块由 2 个分支组成,其中空域分支利用 2D 卷积从单个视频帧中提取空间特征,时域分支利用 3D 卷积从帧序列中提取时序特征,2 个分支的输出经过拼接、降维和非线性激活之后输入到下一个 SMART 模块中. 为了更好地建模时序关系,SMART 模块将时序分支中卷积操作的线性组合运算替换为乘法运算,旨在解耦帧内空间信息和帧间时序转换信息,从而使两者尽可能相互独立,减少空间特征对时序建模的干扰. 此外,Wang 等人[71]指出常规的 CNN 模型通过局部连接进行卷积和池化运算,在逐层池化的过程中损失了大量的时序依赖信息,导致模型对视频全局特征的建模能力有限. 为此,他们提出了一种基于非局部计算模块的 CNN 模型,旨在弥补局部卷积操作在全局时序建模方面的不足. 其核心思想是借鉴图像去噪领域中的非局部平均(non-local means)算法,在像素级定义一个建模当前位置信号和全局候选信号之间关系的函数,并将该函数嵌入到底层卷积网络中,以构建层次化的非局部特征学习模型.

在视频动作识别问题上,人和场景内物体之间的交互信息对于提升识别精度至关重要. Ma 等人[72]提出建模物体交互(object interactions)的动作识别思路,学习高层次视频语义信息. 该方法首先通过判断多个物体在 3D 空间中的重叠度来表示交互信息,

并设计注意力模块完成对全局视频特征的提取.同时,对单独的视频帧生成多个候选区域,输入到LSTM网络中实现局部交互特征提取,有效提高了细粒度动作识别的性能.

虽然有监督视频分类已经取得了长足的进展,但有监督模型的训练依赖大量标注数据,而且在有限视频类别上训练得到的有监督模型难以扩展到新增类别上.为了克服这种限制,零样本视频分类逐渐成为了一个新的研究热点.零样本视频分类旨在通过建立类别标签之间的语义关联,实现从已知类别到未知类别的知识迁移,使得模型能够在没有任何未知类别训练数据的条件下对未知类别视频样本进行分类.

零样本视频分类方法常依赖语义辅助信息(如属性、词向量等)学习一个特征嵌入函数,通过该函数将已知类别和未知类别的视频特征嵌入到一个共享的语义空间中,然后通过相似性度量完成对未知类别的分类.例如,Xu等人[73]通过流形正则化和数据增强方法学习视觉-语义映射函数,在多个视频数据集上得到了较高的零样本分类准确率.Gan等人[74]提出一种多模型保序映射方法,通过构建类别类比池(analogy pool)来衡量不同视频之间的语义联系,提高了类别间知识迁移的效率.不同于上述方法,Zhang等人[75]提出了对抗式特征合成的零样本视频分类方法,利用多粒度语义信息推断和互信息关联约束建模视频特征和文本表示之间的联合分布,用生成式对抗方法合成未知类别的视频特征,将零样本学习转换为有监督问题,借助有监督方法的优势提升零样本视频分类的准确率.

2.2 视频目标检测

目标检测是指计算机自动定位并识别图像视频中的感兴趣目标,是计算机视觉领域的一个重要问题.在过去的几年中,图像目标检测得到了长足的发展,研究者们提出了R-CNN(regions with CNN features)[76],YOLO(you only look once)[77]等一系列模型,从准确率和速度两方面显著提升了图像目标检测的性能.视频目标检测具有更广泛的应用需求,如视频监控、无人驾驶等.然而,视频中存在视角变化、运动模糊、遮挡、形变等问题,给视频目标检测带来了很大挑战.相比于图像目标检测,如何利用视频的时序上下文信息提升视频目标检测的性能,是研究者们关注的焦点.

Kang等人[78]首先利用图像目标检测模型对单个视频帧进行目标检测,然后设计了多上下文抑制策略和运动引导的传播策略,分别利用上下文和运动信息降低误检率和漏检率.同时结合视频跟踪算法优化检测结果的时序一致性.基于该方法,Kang等人在ILSVRC 2015挑战赛的视频目标检测(object detection from video, VID)任务中取得第1名.

Zhu等人提出了端到端的视频目标检测模型[79-80].其中,DFF(deep feature flow)方法[79]关注于提高视频目标检测的速度.该方法把视频帧区分为关键帧和非关键帧,利用卷积神经网络对关键帧进行特征提取.然后利用光流估计和特征传播方法,基于关键帧的特征图计算得到非关键帧的特征图.最后对特征图进行分类得到最终的目标检测结果.由于耗时的卷积特征提取操作只在数量较少的关键帧上进行,而光流估计和特征传播的计算效率远高于卷积特征提取,因此DFF提高了检测效率.FGFA(flow-guided feature aggregation)方法[80]则关注于提高视频目标检测的准确率.该方法对每个视频帧都利用卷积神经网络进行特征提取,然后结合光流估计方法自适应聚合前后帧的特征图以提高当前帧的特征图质量,提高了目标检测的准确率.而Zhu等人后续的工作[81]则结合了上述2个方法的优点.该方法在DFF和FGFA的基础上提出了3个策略:1)只在数量较少的关键帧上进行特征聚合以提高特征图的质量并减少计算开销;2)对于由特征传播得到的非关键帧的特征图,识别其中质量较差的局部块并进行改善;3)自适应地动态选择关键帧.通过以上3个策略,该方法实现了视频目标检测准确率和速度的平衡.

除了一般通用目标的检测之外,很多研究还针对一些特定目标的检测展开,比如标志、行人、文字等.下面分别介绍它们的典型方法和最新进展.在标志检测方面,早期方法主要是基于关键点的局部特征匹配,常见的局部特征包括SIFT(scale-invariance feature transform)[82]等.关键点的误匹配是影响这类方法效果的最大因素,因此研究者们致力于设计描述能力更强的特征[83]以及优化匹配结果[84].深度学习方法的引入也促进了标志检测技术的发展.Yang等人[85]结合手工特征和卷积神经网络,首先基于手工特征获取标志候选框,然后利用卷积神经网络对候选框进行分类,在保证检测速度的同时,也取得了很好的检测准确率.

行人检测一直都是学术界和工业界关注的重要研究问题.Zhu等人提出了一系列方法,基于Boosting检测框架,利用隐含语义特征学习[86]、多任务

学习[87-88]、分组代价敏感方法[89]建立了能够应对遮挡、低分辨率等复杂条件下的行人检测模型. 针对行人检测的多尺度问题,Li 等人[90]提出了 SAF R-CNN (scale-aware fast RCNN)方法,针对大尺度和小尺度行人样本分别训练 2 个子网络,并设计了尺度感知权重策略,根据候选框的行人尺度对 2 个子网络赋予不同的权重. 该方法能够利用多尺度行人样本的不同视觉特征,实现了对多尺度行人检测的鲁棒性. Zhang 等人[91]认为从卷积神经网络的特征图可以提取行人的语义信息,而对于被遮挡的行人,应该只对可见部分进行特征提取,降低被遮挡部分的负面影响. 因此,他们提出在卷积神经网络特征图的不同通道(channel)上应用注意力机制,在学习过程中增大可见部分特征的权重,降低遮挡部分特征的权重,以此提高行人检测的准确率.

视频文字的检测与识别对于视频内容的自动理解具有重要意义,是研究者们关注的重点之一. Yi 等人[92]提出了一种基于彩色聚类的文字检测方法,结合 YUV 和 RGB 颜色空间得到彩色边缘图,并利用近邻传播聚类算法得到多个边缘子图,进而通过边缘子图的投影确定文字区域. Mishra 等人[93]首先通过滑动窗得到字符探测结果,然后通过条件随机场(conditional random field, CRF)模型利用语言先验对探测结果进行验证,从而得到最终的识别结果. Shi 等人[94]提出了基于注意力机制的序列化识别模型,利用编码器-解码器结构进行文字检测和识别.

3 跨媒体检索

随着互联网中图像、视频、文本、音频等多媒体数据的日益增多,用户的信息检索需求也日益增长. 然而,目前常用的信息检索方式还是以单媒体检索为主,例如图像检索、文本检索等. 这些检索方式只能返回与查询数据相同媒体类型的检索结果,限制了信息检索的全面性和灵活性. 因此,用户需要一种跨越不同媒体类型的新型检索方式,能够根据任意媒体类型的查询,检索得到多种媒体类型的结果,将其称之为跨媒体检索[95]. 如图 4 所示,用户任意给定一种媒体类型数据(如一张拍摄的北京大学照片)作为查询,通过跨媒体检索能够自动检索得到与查询主题相关的各种媒体数据,不仅包括有关北京大学的校园图像,也包括北京大学的文本描述、视频介绍、音频资料等,检索方式更加方便、灵活. 由于不同媒体类型数据之间存在相关性、互补性等语义关联关系,相比单一媒体能更加全面准确地进行语义表达. 因此,跨媒体检索能够克服传统单媒体检索信息有限、媒体类型单一的问题,更加符合人脑的多模态感知与认知方式,提升了用户的信息获取效率和检索体验.

由于不同媒体类型数据之间存在“异构鸿沟”,导致图像、文本、音频等不同媒体类型数据的特征表示不一致,无法直接度量它们的相似性,这也成为了

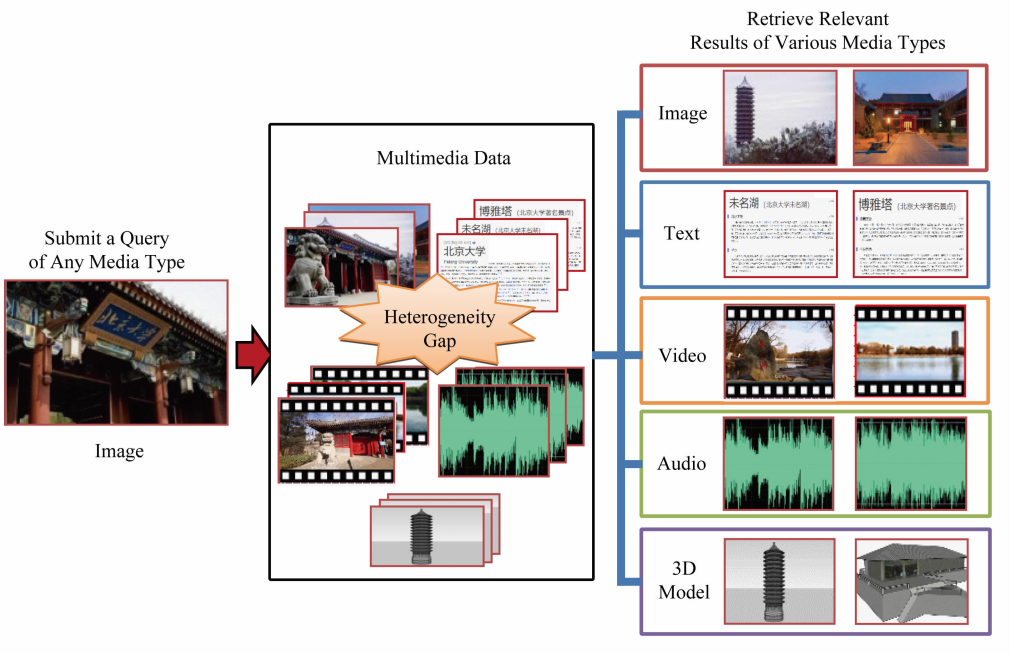

Fig. 4 An example of cross-media retrieval

图 4 跨媒体检索举例示意图

跨媒体检索面临的主要挑战.然而,不同媒体数据虽然在表征上彼此异构,但在语义上却相互关联,这使得跨媒体检索成为可能.例如关于鸟的图像和鸟的叫声,它们的底层特征表示不一致,无法直接度量相似性.但它们都描述了"鸟"这一语义概念,因此在高层语义空间中会彼此接近.现有方法主要通过分析不同媒体数据之间的语义关联来缩短"异构鸿沟",并为不同媒体数据生成同构的统一表征,从而突破"异构鸿沟"造成的相似性不可度量的问题.当不同媒体类型数据都被映射为统一表征之后,即可利用常用的距离度量方法(如欧氏距离)进行语义相似性计算,实现跨媒体检索.针对跨媒体检索问题,Peng等人[95]对跨媒体检索的基本概念、现有方法、主要挑战进行了综述,并构建了国际上数据量最大、媒体类型最多的跨媒体检索数据集 PKU XMediaNet①.这里主要介绍3类方法:跨媒体检索的传统方法、深度学习方法和 Hash 方法.

3.1 跨媒体检索的传统方法

传统方法主要通过统计分析的方式学习映射矩阵,其中一个代表性方法是典型相关分析[96](canonical correlation analysis, CCA).它通过分析2种媒体特征之间的成对关联关系,学习一个能够最大化成对相关性的共同空间,将不同媒体的特征映射到这个共同空间得到相同维度的向量表示,实现跨媒体统一表征.CCA 在跨媒体检索中应用广泛,后续很多工作都是在 CCA 方法基础上的扩展.如 Rasiwasia 等人[97]将语义类别信息与 CCA 结合,通过 CCA 方法得到图像和文本的共同空间,然后使用 logistic 回归实现语义概念的分类;Gong 等人[98]提出了多视角典型相关分析方法(multi-view CCA),在对2种媒体特征进行相关性分析的同时,将语义标签作为第3种视角进行学习;Ranjan 等人[99]提出了多标签的典型相关分析方法(multi-label CCA),能够支持多标签跨媒体数据的关联分析.除了 CCA 之外,跨模态因子分析方法[100]通过最小化成对数据在共同空间中的 Frobenius 范数学习统一表征,在跨媒体检索中取得了比 CCA 方法更好的效果.

上述方法往往只能进行2种媒体的成对关联学习,无法支持更多媒体类型的同时交叉检索.为突破这一限制,一些工作引入了图规约方法,利用图模型在关联表达上的灵活性来学习多种媒体类型数据的统一表征.Zhai 等人[101]提出了统一图规约的异构度量学习方法,利用共同空间中的数据表征构建联合图规约项,但只针对2种媒体类型的数据生成统一表征.他们的后续工作[102]提出了统一表征学习方法,能够为不同媒体类型数据分别构建图模型,联合挖掘不同媒体类型数据的关联关系及高层语义信息,首次将统一表征的媒体类型从2种提升到5种,包括图像、文本、视频、音频和3D图形.进一步,Peng 等人[103]构建了统一的跨媒体关联超图,充分挖掘不同媒体数据的全局信息和与细粒度信息,并结合半监督规约来学习更加准确的跨媒体统一表征.

除了上述工作,还有很多其他方法被提出,例如:基于排序学习的方法[104-105]通过分析不同媒体类型数据的排序信息,实现跨媒体相关性排序;基于字典学习的方法[106]将数据分解为字典矩阵和稀疏系数2个部分,通过不同媒体稀疏系数的相互转换实现跨媒体检索.

3.2 跨媒体检索的深度学习方法

近年来,深度学习在多媒体领域取得了巨大进展,启发了一系列基于深度学习的跨媒体检索方法.这些方法旨在利用深度神经网络对非线性关系的抽象能力,促进跨媒体关联分析和统一表征学习.现有方法一般通过构建多路网络结构建模不同媒体类型数据之间的关联关系.Ngiam 等人[107]扩展限制玻尔兹曼机模型(restricted Boltzmann machine, RBM),提出了双模态深度自编码器.在该方法中,2路网络通过一个共享编码层相连,同时在顶层建模不同媒体数据的重构信息,这样在保持编码可重构性的同时学习跨媒体关联,以此生成跨媒体统一表征.后续的很多工作都受此启发,如 Srivastava 和 Salakhutdinov 提出了基于深度玻尔兹曼机模型(deep Boltzmann machines, DBM)的多模态深度玻尔兹曼机(multimodal DBM)[108],该模型通过2个 DBM 模型分别建模图像和文本数据,且它们可以相互影响,通过整个网络的联合训练得到更好的统一表征.Andrew 等人[109]提出了深度典型相关分析方法(deep CCA),将传统的典型相关分析方法与深度网络结合,在2个子网络顶层学习不同媒体类型数据之间的关联关系,从而得到2种媒体类型的特征表示到共同空间的非线性变换.Feng 等人[110]提出了关联自编码器方法,通过中间层连接2路子网络,在统一表征学习的过程中能够同时考虑不同媒体类型数据的关联和

① http://www.icst.pku.edu.cn/mipl/XmediaNet

重构信息. Wang 等人[111]首先将栈式自编码器应用到跨媒体检索中,利用一个成对的子网络结构来建模跨媒体关联信息. Peng 等人[112]提出了跨媒体多深度网络结构,同时考虑媒体内和媒体间的关联信息,通过层叠式的学习策略得到统一表征. 他们进一步提出了跨模态关联学习方法[113],同时建模不同媒体类型数据的粗细粒度信息,并结合多任务学习自适应地平衡媒体内语义类别约束和媒体间成对关联约束的学习,提高了跨媒体检索的准确率.

上述工作主要以传统手工特征作为输入进行统一表征学习,而后续工作如深度语义匹配方法[114]则利用卷积神经网路(CNN),直接以原始图像像素作为输入,学习具有更强表示能力的跨媒体统一表征. Peng 等人[115]针对不同媒体类型之间不平衡的关联信息展开研究,提出了基于特定模态语义空间建模的跨模态相似性学习方法. 该方法通过基于注意力机制的联合关联学习实现不同媒体数据的相互映射,最后使用动态融合的方法进一步挖掘不同媒体语义空间的互补性,提高了跨媒体检索的准确率. Qi 等人[116]提出了跨模态双向翻译方法,将机器翻译的思想应用到跨媒体检索中,同时结合强化学习来提升跨媒体关联学习的准确率. 他们还提出了跨媒体关系注意力网络[117],实现了不同媒体数据之间全局、对象、对象关系 3 个级别的对齐,同时通过三者的融合有效地促进了跨媒体关联学习. Wang 等人[118]提出了对抗式跨媒体检索方法,通过对抗学习的策略挖掘跨媒体关联关系. Gu 等人[119]将生成模型与传统跨媒体深度特征学习框架相结合,在学习全局特征的基础上,利用图像、文本的相互生成充分挖掘不同媒体数据的局部信息,促进了不同媒体数据之间的关联学习. Fan 等人[120]提出了多感知融合网络,通过为图像生成额外的文本描述,在弥补训练数据不足的同时,充分学习不同媒体数据之间的语义关联,提高了检索的准确率.

此外,基于深度学习的方法往往依赖于大量的标注数据,才能获得较好的模型训练和检索效果. 但跨媒体数据的标注成本巨大,需要同时涉及图像、文本、视频、音频、图形等多种媒体数据的标注. 因此如何利用已有单一媒体的训练数据支持跨媒体检索模型的训练,就成为了一个重要的问题. Huang 等人[121-122]将迁移学习应用到跨媒体关联学习中,提出了跨模态混合迁移网络方法. 该方法的网络结构如图 5 所示,仅使用包含单一媒体类型的源域数据,同时进行媒体内与媒体间的知识迁移,促进目标域的跨媒体模型训练,缓解了跨媒体数据标注成本巨大导致模型训练困难、检索准确率低的问题.

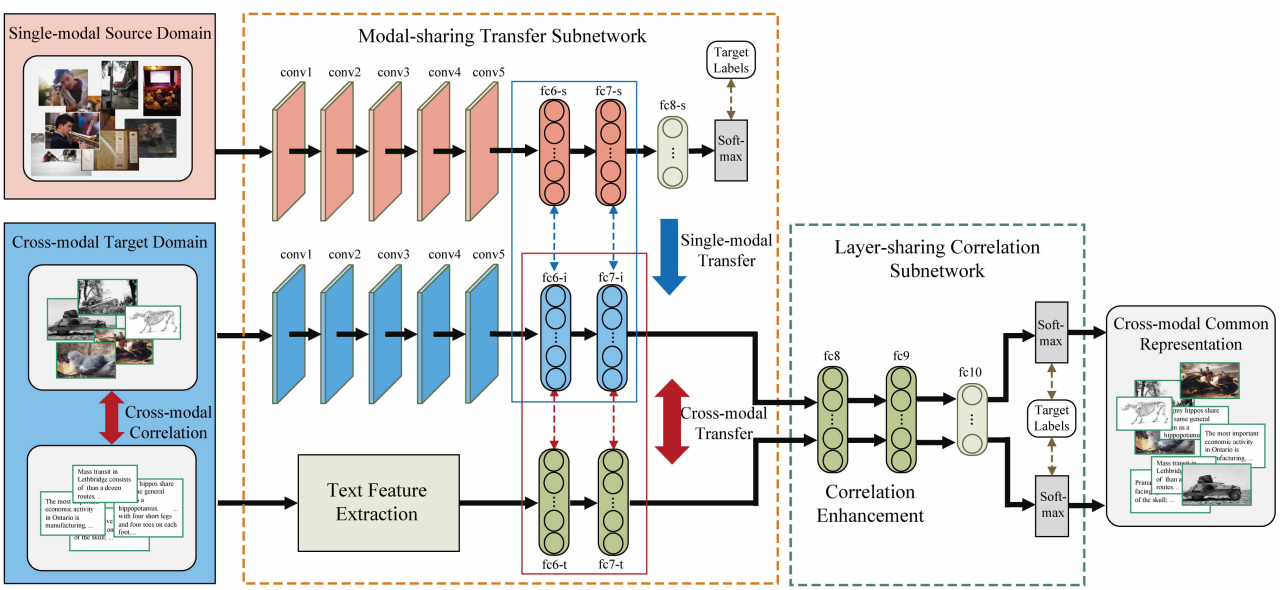

Fig. 5 The overall framework of cross-modal hybrid transfer network[121]

图 5 跨模态混合迁移网络结构示意图[121]

与传统方法相比,基于深度学习的跨媒体检索方法能够充分利用深度神经网络的非线性建模能力以及逐层抽象机制,有效提升对复杂跨媒体关联的分析能力,提高跨媒体统一表征的检索准确率.

3.3 跨媒体检索的 Hash 方法

随着多媒体数据总量的不断增长,跨媒体检索的速度成为了影响实际应用效果的一个关键因素. 在单媒体检索中,Hash 方法是加速检索的有效手段

之一,能够将高维特征转换为二进制的 Hash 编码,从而利用高效的汉明距离度量提高检索速度.与单媒体 Hash(如图像 Hash)不同,跨媒体 Hash 需要同时将多种媒体类型的数据映射到一个共同的汉明空间中并生成 Hash 编码.现有的跨媒体 Hash 方法可以分为无监督方法、有监督方法和深度方法.

无监督跨媒体 Hash 方法的主要思想是通过学习 Hash 映射函数,在不使用语义类别标注的情况下,将不同媒体数据映射到一个共同的汉明空间中,并在该空间内维持原有不同媒体数据之间成对的关联信息.例如 Song 等人[123]以媒体内和媒体间的一致性关联作为约束,为不同媒体数据学习共同的汉明空间.Long 等人[124]通过学习关联最大化映射,将不同媒体数据映射到一个同构空间,并使用量化的方式将同构空间中的媒体特征转化为 Hash 码,实现跨媒体 Hash 检索.Liu 等人[125]提出融合相似 Hash,认为将实体的不同模态特征作为整体计算得到的融合相似性能够利用模态间的互补性,并提出基于图的融合相似性计算方法,以此指导 Hash 函数的学习.

有监督的跨媒体 Hash 方法利用标注好的语义类别标签来指导 Hash 函数学习,往往能取得比无监督方法更好的检索效果.Bronstein 等人[126]提出了跨模态相似性敏感 Hash 方法,利用 boosting 方法学习 Hash 函数.Wei 等人[127]提出了异构转换 Hash 方法,通过学习转换器将不同汉明空间中的跨媒体数据进行对齐,从而实现跨媒体 Hash 检索.Lin 等人[128]将语义信息矩阵转化为概率的分布,通过最小化语义概率分布和生成汉明空间分布间的 KL 散度,实现 Hash 码的学习.

此外,近年来基于深度学习的跨媒体 Hash 方法也得到了广泛的关注.如 Zhuang 等人[129]利用神经网络的抽象学习能力,通过维持媒体内的差异性以及媒体间的成对关联信息实现 Hash 函数学习.Ye 等人[130]考虑到不同尺度特征蕴含大量的互补信息,以及尺度特征之间存在丰富的关联信息,提出序列化多尺度 Hash 方法,在同时建模多尺度特征的同时,实现尺度间的关联挖掘,进一步提高了检索效果.Zhang 等人[131]借鉴了对抗式学习的思想,提出了半监督跨媒体生成式对抗 Hash 方法,利用生成式对抗网络(generative adversarial networks, GANs)[132]寻找无标注数据中容易混淆的训练样本,增强模型对易混淆样本的识别能力,从而提高跨模态检索效果.Zhang 等人[133]进一步提出了无监督跨媒体生成式对抗 Hash 方法,借助关联图模型实现了无监督下的模型对抗训练,有效地建模数据流形结构用于 Hash 码生成,并取得了无监督条件下更好的检索结果.

4 视觉描述与生成

视觉描述与生成是一种视觉内容高层语义理解任务,旨在实现对视觉内容的高层语义认知与自然表达,涉及计算机视觉、自然语言理解、机器学习等多个研究领域.视觉描述与生成包含用自然语言描述图像视频的视觉内容,以及根据自然语言描述生成图像视频等多个研究方向,具有重要的潜在应用价值,近年来备受学术界和工业界的关注.本节将重点介绍图像视频的文本描述生成、文本到图像生成、文本到视频生成.

4.1 图像视频的文本描述生成

图像视频的文本描述生成(image/video captioning)是指计算机自动生成对图像和视频内容的自然语言描述,在人机交互、帮助视障人员理解图像视频内容方面具有潜在的应用价值.

早期图像视频的文本描述生成方法主要包括基于模板的方法和基于检索的方法.基于模板的方法首先检测图像视频中的对象、属性、概念以及对象关系等内容,然后利用预定义的语言模板,将检测到的视觉内容和语句的组成部分(例如 subject,verb 和 object)进行对齐,以此生成文本描述.Kulkarni 等人[134]首先检测图像中的对象,并预测对象的属性和对象间的介词关系,然后构建 CRF 模型预测三元组形式的标签信息,最后根据语言模板生成语句.Yang 等人[135]将语句的核心结构表示为"名词-动词-场景-介词"四元组的形式,并利用 HMM(hidden Markov model)模型选择最合适的四元组来生成语句.这个方法还结合基于大规模语料训练的语言模型来提高名词、动词、场景和介词的预测准确率.基于模板的方法依赖对象、属性等检测的质量,而且其语句生成过程依赖预定义的语言模板,导致生成的语句结构比较单一,多样性受限.基于检索的方法[136]采用信息检索的模式来"生成"语句,即从人工构建的语句集合中检索出与图像语义相似的语句,并根据检索得到的语句生成最终的语句描述.虽然这类方法能够得到与人工描述密切相符的语句,但是所得到的语句受限于人工构建的语句集合,并且语句集合不易扩展.

近年来,基于 RNN 的序列学习模型在机器翻译领域取得了极大的进展.受此启发,研究人员将图像视频的文本描述生成看成是一个“翻译”过程,构建编码器-解码器(encoder-decoder)模型,首先将图像视频的视觉内容编码成特征向量,然后利用 RNN 模型将特征向量解码为文本描述.这类方法通过对视觉内容和文本序列进行联合建模,直接从视觉内容中生成文本描述,不依赖具体的语言模板,因而能够生成语法结构灵活、更加符合人类语言表达习惯的语句.目前这类方法已经成为图像视频的文本描述生成任务的主流方法.编码器-解码器模型中的编码器通常由 CNN 构成,对于视频而言,除了常规的 2D CNN 网络以外,C3D 等 3D CNN 以及 RNN 也通常用于编码视频的时序信息.解码器通常由 RNN 构成,LSTM 等是常用的解码器模型.

Vinyals 等人[137]提出 NIC(neural image caption)模型,率先将编码器-解码器模型用于图像的文本描述生成.该模型利用预训练的 GoogLeNet 网络作为编码器,利用 LSTM 网络作为解码器生成文本描述,在多个数据集上取得了令人鼓舞的结果.Venugopalan 等人[138]则将上述方法直接扩展到视频的文本描述生成任务,利用 CNN 提取单个视频帧的特征,并采用平均池化的方法得到整个视频的特征表示,然后通过 LSTM 得到文本描述.但该方法忽略了视频帧的时序特点.随后,他们进一步提出 Seq2Seq(sequence to sequence)模型[139],将编码器和解码器全部构建成序列学习模型.具体地,利用编码端的 LSTM 建模视频帧序列的时序信息,编码得到视频的高层语义表示,然后利用该语义表示特征向量,初始化解码端 LSTM 的隐层状态,进而生成文本描述.后续工作通常采用上述编码器-解码器结构作为基本框架,通过构建多种注意力机制,结合多模态特征以及引用外部知识等方式提高图像视频的文本描述生成效果.

Xu 等人[140]将注意力机制引入 NIC 模型中以生成图像的文本描述.该方法利用 CNN 的卷积层提取多个图像块的特征,然后在 LSTM 网络中引入软注意力(soft attention)和硬注意力(hard attention)机制.软注意力机制为每个图像块学习一个概率,表示该图像块的显著程度,最后通过加权的方式将多个图像块的特征表示进行融合.而硬注意力机制则通过采样的方式直接选择最显著的图像块,该机制通过强化学习进行优化.该方法使用上述 2 种注意力机制选择表示图像显著性信息的视觉特征,进而通过 LSTM 生成更准确的文本描述.Zhang 等人[141]提出一种注意力引导的层次化对齐方法,通过建模视觉内容和文本描述之间多层次的语义一致性信息,提高视频的文本描述生成效果.Hori 等人[142]提出利用视频的多模态信息生成文本描述.该方法分别提取视频的帧特征、时空特征以及音频特征,并设计了基于注意力机制的 LSTM 对多模态特征进行融合.Venugopalan 等人[143]结合在大规模标注数据上训练好的目标检测模型以及从外部语料库提取的分布式语义信息,对图像中的“新”对象(novel object,即未在训练集中出现的对象)进行识别和描述.另外,针对 LSTM 模型结构复杂、训练速度慢、内存开销大的问题,Aneja 等人[144]构建了基于卷积模型的解码器,并在图像的文本描述生成任务上取得了与 LSTM 解码器相近的结果.

上述方法存在 2 个问题:1)训练阶段使用交叉熵损失(cross entropy loss)函数,通过最大化后验概率进行训练,但测试阶段使用 BLEU,METEOR 和 CIDER 等评价指标,损失函数和评价指标不统一.2)RNN 解码器在训练阶段使用真实文本描述的单词作为每个时刻的输入,但在测试阶段使用上一时刻预测得到的单词作为下一时刻的输入,这导致了 exposure bias 问题[145].研究者们引入强化学习方法,通过直接优化 BLEU 等评价指标以及在训练过程中使用采样策略来解决上述 2 个问题.Liu 等人[146]提出了基于强化学习的图像文本描述生成方法,该方法利用策略梯度(policy gradient)方法训练解码器,并且利用 BLEU,METEOR,CIDER 和 SPICE 等评价指标的组合设计奖励(reward)函数.Pasunuru 等人[147]提出基于强化学习的视频文本描述生成方法,该方法在训练阶段联合使用强化学习损失函数以及交叉熵损失函数,前者约束生成的文本描述与评价指标相适应,后者保证了文本描述的可读性和流畅性.

一般的图像视频的文本描述生成任务是为输入的图像或短视频生成一句文本描述,通常是一句英文语句.在此之外,一些工作还探索了其他形式的图像视频的文本描述生成任务.Johnson 等人[148]关注于 dense captioning 任务,旨在为图像中的多个区域分别生成文本描述,包含区域定位和文本描述生成 2 个过程.Krishna 等人[149]关注于 dense-captioning events 任务,旨在为长视频中的多个事件分别生成文本描述,包含事件的时序定位和文本描述生成 2 个过程.Yu 等人[150]提出了一种细粒度的视频文本

描述生成(fine-grained video captioning)任务,对运动视频中多个人物的具体动作进行描述.例如对于一个“打篮球”的视频,一般的视频文本描述生成任务只是宏观地描述“打篮球”这个事件,而细粒度的视频文本描述生成任务则会描述“传球”、“运球”、“灌篮”等细粒度的具体动作.一些研究者还关注于多语言/跨语言的文本描述生成任务,旨在研究生成非英语语言的文本描述.例如,Lan 等人[151]提出跨语言的文本描述生成模型,基于英语的文本描述数据集以及机器翻译方法,为图像生成汉语语言的文本描述.

4.2 文本到图像生成

人们通常会采用检索的方式来寻找有用的信息,例如文本检索、图像检索等.然而,传统检索方式只能为用户提供数据库中已有的数据,并且需要进行大量的人工标注进行识别模型训练,限制了信息获取的灵活性.文本到图像生成是指,用户提供一段文本描述,计算机能够自动生成符合这段文本描述内容的图像.文本到图像生成大大提高了图像信息获取的灵活性和全面性,可以用于多个重要领域,如公安领域的模拟画像、教育领域的概念启蒙、艺术领域的视觉创作等.

随着近年来变分自编码器(variational autoencoder, VAE)[152]与 GAN 等深度生成模型的兴起,文本到图像生成取得了一系列进展.Yan 等人[153]利用 VAE 实现了一种文本到图像生成的方法,可以利用视觉属性生成图像.该方法的作者认为图像是前景和背景的组合,所以提出了分层生成式模型.该生成模型拥有分组的潜变量,可以通过变分自编码器进行端到端训练,从而具备文本到图像生成的能力.

Reed 等人[154]以条件生成式对抗网络(conditional generative adversarial networks, conditional GANs)为基础,提出了 GAN-INT-CLS 方法,能够根据文本生成视觉效果更加真实且符合语义描述的图像.他们首先提出了一种可以表达文本中视觉信息的特征,然后将这种特征作为输入,利用生成网络生成一幅图像,再利用判别网络对该图像进行解析,判断该生成图像与输入文本的关联性与真实性.由于生成网络希望生成的图像能够“以假乱真”,判别网络希望可以区分生成图像与真实图像,两者形成对抗式训练过程,互相促进、不断提高,最终使得生成网络具有文本到图像的生成能力.Reed 等人[155]提出了一种“内容-位置”生成式对抗网络 GAWWN,通过给出的“内容-位置”的说明来生成图像.GAWWN 方法将空间遮挡和裁剪模块合并到文本的生成式对抗网络中,同时用一组归一化坐标表示部件位置作为条件,使得生成网络和判别网络能够使用乘法门控机制来关注相关部件的位置,从而可以为输入文本生成空间结构上更加合理的图像.

Zhang 等人[156]借鉴了将 2 个生成式对抗网络叠加在一起的结构[157],并采用两阶段训练方法分别训练 2 个生成式对抗网络,实现了较大尺寸图像的生成.2 个阶段的生成式对抗网络的作用是:阶段 1 的生成网络利用文本描述生成包括物体主要的形状和颜色的低分辨率图像;阶段 2 的生成网络将阶段 1 的结果和文本描述作为输入,生成细节丰富的高分辨率图像.同样以生成细节丰富的高分辨率图像为目标,Zhang 等人[158]利用生成网络的层次结构引入了分层嵌套对抗目标函数,从而规范了生成图像的中层特征表达并协助生成网络拟合真实图像的数据分布.该方法提出了一种生成网络结构,以更好地适应判别网络并将生成的低分辨率图像扩展到高分辨率图像.另外该方法采用一种多用途对抗性损失函数来促进模型进一步挖掘图像和文本中的潜在关联信息,进一步提高了生成图像的视觉质量以及生成图像与文本的内容一致性.

Xu 等人[159]在生成式对抗网络的基础上,引入注意力驱动模型和多阶段精化模型,提出了注意力生成式对抗网络(attentional generative adversarial network, AttnGAN)来应对细粒度的文本到图像生成问题.通过注意力驱动模型,AttnGAN 可以通过关注自然语言描述中的相关单词来合成图像的不同子区域的细粒度细节;同时多阶段精化模型可以迭代地提高生成图像的视觉质量,最终生成接近真实的图像.此外,该方法提出一种深度注意力多模态相似度损失函数,可以在细粒度层面保证生成图像与输入文本之间的内容一致性.

不同于基于 VAE 和 GAN 的方法,Yuan 等人[160]提出了对称蒸馏网络(symmetrical distillation networks, SDNs).该网络的结构如图 6 所示,由一个源判别模型和一个目标生成模型组成,两者具有对称的结构,可以将源判别模型(例如在 ImageNet 数据集上预训练的 VGG19 模型)的知识蒸馏到目标生成模型中.具体地,SDN 提出了一种新的两阶段蒸馏方法,其中阶段 1 是直接蒸馏,主要使生成模型学习物体的基础形状和颜色;阶段 2 是间接蒸馏,主要使生成模型从细节上学习物体的形状和颜色.

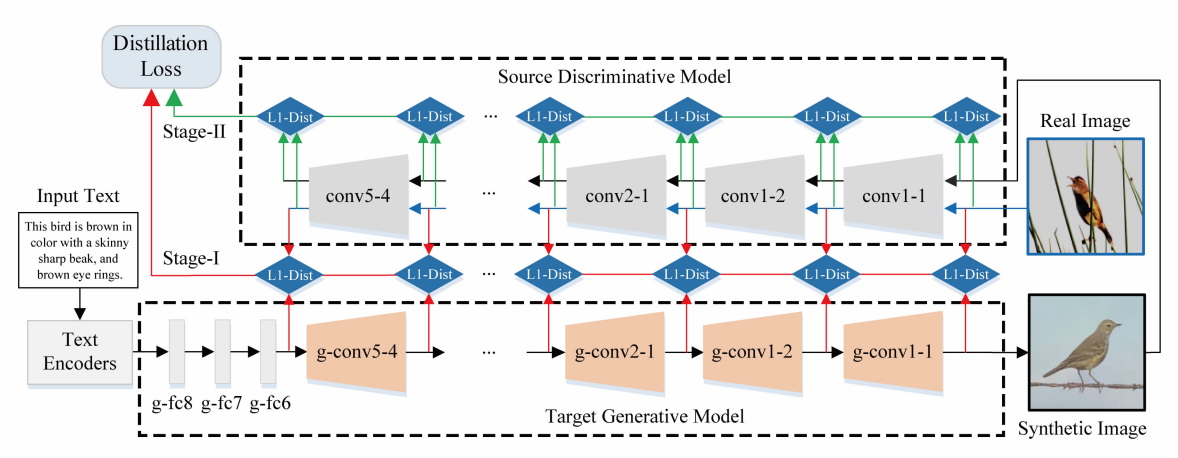

Fig. 6 The architecture of symmetrical distillation networks (SDNs)[160]

图 6 对称蒸馏网络框架[160]

这种对称网络结构和两阶段蒸馏方法将通用判别模型的知识蒸馏到文本生成图像模型中,从而使得生成图像在内容上更加符合输入文本的描述,在数据分布上也更加接近真实图像.

由于以上方法在输入文本包含多个视觉目标的情况下都存在效果不好的问题,Johnson 等人[161]针对这个现象提出了基于场景图(scene graph)的文本到图像生成方法. 该方法可以明确推理多个视觉目标的位置及关系,为包含多个视觉目标的文本生成内容一致且合理的图像. 该方法将场景图作为文本输入,使用图卷积来处理输入的场景图,然后通过预测对象的边界框和分割蒙版计算场景布局,最后利用级联细化网络将布局转换为图像. 该方法所使用的生成网络利用判别网络进行对抗训练,提高生成图像的质量并保证生成图像与文本内容一致.

4.3 文本到视频生成

文本到视频生成旨在根据输入的文本信息,生成符合文本内容且具有时序相关性的连续视频帧. 相比于文本到图像生成,文本到视频生成需要同时考虑文本和视频之间的语义一致性以及视频时空结构的连续性,使得现有的文本到图像生成方法无法直接应用于文本到视频生成.

Mittal 等人[162]利用变分自编码器实现文本到视频的生成,将注意力机制和循环变分自编码器(recurrent-VAE, R-VAE)结合,设计了一种多帧同步的 Sync-DRAW 模型. 该模型由读机制、R-VAE 和写机制 3 部分组成. 其中读机制负责利用循环编码器从视频帧中挑选感兴趣区域(region of interest, RoI),R-VAE 负责从所选择的 RoI 中学习视频的潜在分布,最后由写机制负责生成集中于 RoI 的连续视频帧. 类似地,Marwah 等人[163]提出了一种基于软注意力机制的方法,根据文本描述信息同时对视频中的长、短时上下文进行建模,以增量的方式生成视频. 具体而言,他们采用双向 LSTM 网络对文本信息进行编码,用 Conv-LSTM 完成长时上下文的建模.

除了变分自编码器之外,生成式对抗网络也被用于文本到视频的生成. Pan 等人[164]提出了一种基于条件生成式对抗网络的文本到视频生成框架 TGANs-C(temporal GANs conditioning on captions). TGANs-C 利用 LSTM 对文本描述进行特征编码,将其与高斯噪声向量拼接之后输入到生成器中,再借助 3D 卷积实现连续视频帧的生成. 该方法设计了 3 种判别器:视频判别器、帧判别器和运动判别器,以实现对视频时序和语义信息的建模.

Li 等人[165]将变分自编码器和生成式对抗网络结合起来,提出了一种混合式的文本到视频生成框架. 首先利用条件变分自编码器(conditional-VAE)根据文本信息对视频主体特征进行建模,生成一张表示视频背景颜色、物体轮廓信息的图像,然后将该图像和文本同时作为条件,借助条件生成对抗网络(conditional GANs)实现最终的视频生成.

5 视觉问答

视觉问答(visual question answering, VQA)是近年来多媒体内容理解的一个热点问题. VQA 是指根据给定的图像[166]或视频[167]形式的视觉信息,用户提出自然语言形式的问题,由计算机自动生成自然语言形式的答案作为输出. 这是一种极富挑战性的多媒体高层语义理解问题,且往往需要一定的常识和背景知识进行推理.

如图 7 所示,VQA 中的文本问题具有自由开放的特点,可能涉及到对象的种类、位置、属性、关系等. 如图 7(b)的例子,为了回答“你能在这里停车吗?”这个问题,计算机需要理解该问题的意图和图像的内容,同时也需要结合常识判断(停车需要空间,且斑马线不能停车),从而得到正确的答案:不能停车. 作为一种新型的人机交互方式,高效的 VQA 方法不但能提高计算机视觉系统的智能程度(即“视觉图灵测试”[168]),而且具有广泛的潜在应用价值,如语音助手、智能图像检索、视障人士辅助等.

Fig. 7 Examples of visual question answering (VQA)

图 7 视觉问答(VQA)示例

由于 VQA 问题具备重要的研究与应用价值,近年来研究者们提出了很多方法,并取得了一系列研究进展[169-170]. 基于深度学习在图像分类、目标检测、文本生成等问题上的显著进展,现有的 VQA 方法主要以深度网络为基本模型. Antol 等人[166]构建了一个大规模的 VQA 数据集,共包含超过 25 万张图像、76 万个文本问题和 990 万个回答. 他们同时提出了一个基本的 VQA 深度模型(deeper LSTM Q+norm I),包含一个图像分支 VGGNet 和一个文本分支 LSTM. 这 2 个分支的目标是分别得到图像和文本问题的特征表示,它们以点乘的方式进行组合,并通过一个分类器得到答案输出. 该模型虽然较为简单,但体现了深度网络在 VQA 上的效果,并启发了后续的一系列研究工作.

对于 VQA 来说,图像和文本的细粒度语义对齐非常重要. 由于 VQA 中的文本问题往往聚焦于一个或多个对象的属性、关系等,需要对不同局部视觉区域的重要度进行区分,因此往往需要注意力机制的指导. 与单独的图像注意力学习不同,VQA 中的注意力学习需要在基于输入问题的条件下进行. 如 Lu 等人[171]提出了 Co-attention 模型,首先将图像和文本问题分别分割为若干个图像块和文本单元,之后进行图像和文本问题的注意力学习,目标是为图像块和文本单元赋予不同的权重,这样在生成答案时能够强调图像和文本问题对应的关键部分. 在这个过程中,图像的注意力学习以文本表示为指导,文本的注意力学习也以图像表示为指导,形成了图像-文本的注意力交替学习机制. 与 Co-attention 类似,一系列 VQA 方法都引入了注意力学习机制[172],实现了图像与文本问题相关区域的语义对齐,提高了回答的准确率. 此外,还有一些工作[173]关注于图像中对象间的关系推理. 如图 7(d)的例子,需要对“在……之前”这一位置关系进行正确的理解,才能准确回答所提出的文本问题. 然而,现有的视觉关系推理研究主要集中于几何体的形状、大小、位置、颜色等信息,对于现实世界中更加复杂的关系推理则涉及较少.

由于 VQA 中的文本问题具有自由开放的特点,在很多情况下,单纯依靠从图像中得到的信息是不够的,需要引入常识和背景知识进行推理才能正确回答. 图 8 展示了对同一张图像可能提出的 3 个文本问题及其答案. 问题 1 可以从图像中直接得到答案,无需外部知识;而对于问题 2,人可以很容易地给出正确答案,是因为人可以基于“斑马和动物学相关”这一常识做出判断,而这样的信息无法从图像内容本身得到;对于问题 3,则需要更多关于动物的背景知识才能给出正确答案. 针对这类依赖外部知识的情况,Wu 等人[174]提出基于外部知识库的 VQA 方法,首先对图像提取属性作为第 1 组输入,然后基于属性生成图像的文本描述作为第 2 组输入,再根据提取的属性在外部知识库中抽取相关内容作为第 3 组输入,3 组输入合成为一个特征表示,输入到 LSTM 模型中生成问题对应的答案.

Su 等人[175]提出了 VKMN 方法,首先从已有知识库构建一个大规模知识图谱,以 RDF 三元组的形式存储知识. 之后基于记忆网络(memory network)

Q1: How many zebras are there in the image?
A1: Two.
(Directly from the image)

Q2: Is this image related to zoology?
A2: Yes.
(Need common sense knowledge)

Q3: What are the common properties between the animal in this image and giraffe?
A3: Megafauna of Africa; Herbivorous animals.
(Need external knowledge)

Fig. 8 Three question-answer pairs for one image

图 8 对于同一张图像的 3 组问答示例

结构,把 VQA 转化为从知识图谱中检索挖掘答案的过程,实现了外部知识的利用. Wang 等人[176]提出基于外部知识库推理的 Ahab 方法,对图像提取视觉概念构成 RDF 三元组结构,然后与外部知识库中的实体进行关联. 这样将自然语言描述的问题转换为知识库查询模板的形式,把在知识库中执行查询的结果作为答案. Aditya 等人[177]提出了一种基于推理引擎的 VQA 方法,首先从图像和文本问题中分别提取若干个 RDF 三元组表示,将这些三元组与现有知识库中的短语相似性信息一同输入到推理引擎中. 之后结合定义好的若干条答案生成规则,实现了可解释的基于外部知识的 VQA. 这些方法都尝试引入了外部知识,在涉及常识推理的复杂 VQA 问题上取得了准确度的提升. 然而,现有方法通常依赖于预先定义的规则模板,且受到知识库大小等因素的限制.

6 问题与挑战

随着人工智能、互联网等技术的快速发展,图像、视频等多媒体数据在人类社会、物理空间和信息空间相互融合,具有跨模态、跨数据源、跨空间等特性. 尽管多媒体内容理解的研究已经有了很大进展,但仍面临着巨大的需求与挑战,主要列举如下:

1) 跨媒体推理. 现有多媒体内容理解方法主要是以数据驱动的深度学习为主,虽然在图像视频的识别、检索上取得了很大进展,但与人类的智能还相距甚远,无法像人一样具备学习与思考的能力. 原因在于现有方法主要是针对视觉问题,即研究“看”的问题,如同人类的眼睛. 然而,对于感知与认知世界来说,人类大脑的推理能力至关重要. 早期的知识推理方法以文本为主,基于命题和规则在充分定义的前提下进行推理. 但人脑对客观世界的感知和认知,来源于视觉、语言、听觉等多种模态信息. 如何实现跨越不同媒体数据的推理机制以研究“思考”问题,成为了类人智能研究的关键问题. 为此,需要研究数据-知识协同驱动的跨媒体智能分析与推理方法,在跨媒体知识获取、跨媒体知识表征与迁移、跨媒体知识演化上取得突破,建立基于知识逻辑的定向推理和一般性推理机制,实现基于语义理解的跨媒体推理.

2) 小样本训练与学习. 现有方法一般依赖大量的标注样本进行模型训练,才能在识别、检索等任务上取得较高的准确率. 然而,人工收集与标注数据需要耗费巨大的人力物力,以跨媒体检索为例,为了标注“北京大学”这个概念的跨媒体数据,需要标注人员看图像、读文本、看视频、听音频,而现实世界的语义概念种类繁多,人工标注的方式无法处理. 因此,小样本训练与学习就成为了一个重要问题,例如:如何通过数据增广方法[27],基于小样本自动生成大量新训练样本,突破训练样本数量的限制?如何使用跨媒体迁移学习的思想[121],将已有单媒体标注数据如 ImageNet 的知识迁移到跨媒体训练数据中,支持小样本条件下的跨媒体模型训练和学习?

3) 无监督条件下的多媒体内容理解. 面对海量的多媒体数据,需要在小样本学习的基础上进一步减少对标注数据的依赖,实现无监督条件下的图像视频细分类、跨媒体检索等应用,因为用户的分类需求或查询样例多种多样,无法在监督条件下提前训练好. 而人类往往能够自主学习新的知识并发现新的规律与模式. 因此,如何充分利用先验知识以及知识图谱等外部知识库,指导无监督条件下的多媒体内容理解,突破“异构鸿沟”实现跨媒体数据的统一感知与认知,是未来研究面临的重要挑战,也是多媒体内容理解走向实际应用的重要基础.

4) 跨媒体知识图谱. 知识图谱能够描述客观世界中存在的各种实体和概念,以及它们之间的复杂关系,在搜索引擎、知识表示、认知推理等应用中发挥着重要作用. 然而,现有的知识图谱以构建文本实体概念之间的关系为核心. 在多媒体内容理解中,需要拓展传统基于文本的知识体系,通过从图像、视频、

文本等多种媒体数据中抽取知识,形成跨媒体知识图谱并进行跨媒体关联知识表达,实现对多媒体数据高层语义关联的高效计算和综合推理.为实现跨媒体知识图谱的构建与有效利用,应研究的问题包括:如何扩展现有知识图谱的结构定义,支持跨媒体实体、概念和属性等信息及其关联关系的表达?在新增跨媒体数据不断加入的过程中,如何实现跨媒体知识图谱的动态更新机制?如何基于已构建的跨媒体知识图谱进行综合推理?

5) 跨媒体数据相互生成.现有研究主要包括图像/视频的文本描述生成和文本到图像/视频的自动生成2个方面.图像/视频的视觉特征与文本特征之间存在"异构鸿沟",使得视觉内容和文本描述之间的相互生成面临挑战.虽然图像/视频的文本描述生成已经取得了较大进展,但是生成的文本描述与人类理解之间仍存在较大差距.如何生成细粒度、符合人类表达方式的文本描述仍然是一个亟待解决的问题.而文本到图像/视频的自动生成方面的研究起步不久,现有方法所生成的图像/视频在视觉真实性上仍有明显不足,如何自动生成"真实"的图像和视频是下一步研究需要解决的重要问题.此外,在视频、图像、文本、音频等多种媒体内容之间的跨媒体生成方面,如图像生成音频、音频生成视频等还少有涉及.考虑到人类对外界感知与认知是基于不同感官信息融合而成的整体性理解,如何达到各种媒体数据之间的相互生成是一项极具挑战的研究任务.

6) 视觉知识嵌入与推理.深度学习的快速发展与应用,使得图像细分类与检索取得了较大进展,但是缺乏视觉推理能力,难以处理实际应用中复杂的图像细分类与检索问题.如何引入先验知识,通过构建知识图谱等方法将细粒度知识嵌入神经网络,使其具有由局部到整体、由属性到语义的视觉推理能力,是图像细分类与检索的下一步研究方向.此外,现有工作多聚焦于图像,随着视频应用的快速发展,视频细分类与检索将具有重要的研究与应用价值.视频相比于图像具有更多的有用信息,同时也存在大量冗余信息,如何降低冗余,充分整合有用信息强化视频细粒度辨识,也是一个重要的研究问题.

7) 多媒体内容理解的实际应用.现有研究提出并发展了如多媒体检索、描述、生成、问答等应用场景.然而,一方面现有方法在准确率上离实际应用还有较大差距,特别是在生成、问答等高层语义理解问题上的研究还存在诸多不足.另一方面,现有的智能系统往往依赖特定的领域知识,难以满足海量多媒体数据应用的复杂需求.因此有必要研究自主演化的多媒体内容理解技术,形成集底层跨媒体数据表征、索引、关联和高层知识表达、演化、推理等机制为一体,同时具有数据归纳、知识演绎、行为规划能力的跨媒体高效智能计算系统,在智能医疗、智慧城市、智能制造等重要领域发挥重要应用价值.

7 总 结

随着海量的图像、视频、文本、音频等多媒体数据的不断涌现,多媒体内容理解成为了一个研究热点,受到了学术界和工业界的广泛关注.本文对多媒体内容理解的研究现状进行了综述,从图像细分类与检索、视频分类与目标检测、跨媒体检索、视觉描述与生成、视觉问答这5个热点研究方向出发,分别阐述其基本概念与代表性方法.此外,基于上述研究现状,本文进一步阐述了多媒体内容理解所面临的重要挑战,并给出未来的发展趋势,包括跨媒体推理、小样本训练与学习、无监督条件下的多媒体内容理解、跨媒体知识图谱、跨媒体数据相互生成、视觉知识嵌入与推理、多媒体内容理解的实际应用等.本文旨在帮助读者全面了解多媒体内容理解的研究现状,吸引更多研究人员投入相关研究并为他们提供技术参考,推动该领域的进一步发展.

参 考 文 献

[1] McGurk H, MacDonald J. Hearing lips and seeing voices [J]. Nature, 1976, 264(5588): 746-748

[2] Peng Yuxin, Zhu Wenwu, Zhao Yao, et al. Cross-media analysis and reasoning: Advances and directions [J]. Frontiers of Information Technology & Electronic Engineering, 2017, 18(1): 44-57

[3] Zhang Ning, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection [C] //Proc of the 12th European Conf on Computer Vision. Berlin: Springer, 2014: 834-849

[4] Girshick R, Donahue J, Darrell T, et al. Region-based convolutional networks for accurate object detection and segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142-158

[5] Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition [J]. International Journal of Computer Vision, 2013, 104(2): 154-171

[6] Huang Shaoli, Xu Zhe, Tao Dacheng, et al. Part-stacked CNN for fine-grained visual categorization [C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 1173-1182

[7] Krause J, Jin Hailin, Yang Jianchao, et al. Fine-grained recognition without part annotations [C] //Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 5546-5555

[8] Zhou Feng, Lin Yuanqin. Fine-grained image classification by exploring bipartite-graph labels [C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 1124-1133

[9] Xiao Tianjun, Xu Yichong, Yang Kuiyuan, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification [C] //Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 842-850

[10] Zhang Yu, Wei Xiu-Shen, Wu Jianxin, et al. Weakly supervised fine-grained categorization with part-based image representation [J]. IEEE Transactions on Image Processing, 2016, 25(4): 1713-1725

[11] He Xiangteng, Peng Yuxin. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2017: 4075-4081

[12] Zhang Xiaopeng, Xiong Hongkai, Zhou Wengang, et al. Picking deep filter responses for fine-grained image recognition [C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 1134-1142

[13] Zhang Xiaopeng, Xiong Hongkai, Zhou Wengang, et al. Picking neural activations for fine-grained recognition [J]. IEEE Transactions on Multimedia, 2017, 19(12): 2736-2750

[14] Peng Yuxin, He Xiangteng, Zhao Junjie. Object-part attention model for fine-grained image classification [J]. IEEE Transactions on Image Processing, 2018, 27(3): 1487-1500

[15] Fu Jianlong, Zheng Heliang, Mei Tao. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 4438-4446

[16] Wang Yaming, Morariu V I, Davis L S. Learning a discriminative filter bank within a CNN for fine-grained recognition [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4148-4157

[17] He Xiangteng, Peng Yuxin, Zhao Junjie. StackDRL: Stacked deep reinforcement learning for fine-grained visual categorization [C] //Proc of the 27th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2018: 741-747

[18] He Xiangteng, Peng Yuxin, Zhao Junjie. Fine-grained discriminative localization via saliency-guided faster R-CNN [C] //Proc of the 25th ACM Int Conf on Multimedia. New York: ACM, 2017: 627-635

[19] He Xiangteng, Peng Yuxin, Zhao Junjie. Fast fine-grained image classification via weakly supervised discriminative localization [J/OL]. IEEE Transactions on Circuits and Systems for Video Technology. [2018-11-01]. https://ieeexplore.ieee.org/document/8356107. DOI: 10.1109/TCSVT.2018.2834480

[20] Zhang Linbo, Wang Chunheng, Xiao Baihua, et al. Image representation using bag-of-phrases [J]. Acta Automatica Sinica, 2012, 38(1): 46-54

[21] Lin Tsungyu, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition [C] //Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1449-1457

[22] Zhao Junjie, Peng Yuxin. Cost-sensitive deep metric learning for fine-grained image classification [C] //Proc of the 24th Int Conf on MultiMedia Modeling. Berlin: Springer, 2018: 130-141

[23] Liu Xiao, Wang Jiang, Wen Shilei, et al. Localizing by describing: Attribute-guided attention localization for fine-grained recognition [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2017: 4190-4196

[24] Chen Tianshui, Lin Liang, Chen Riquan, et al. Knowledge-embedded representation learning for fine-grained image recognition [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2018: 627-634

[25] Xu Huapeng, Qi Guilin, Li Jingjing, et al. Fine-grained image classification by visual-semantic embedding [C] //Proc of the 27th Int Joint Conf on Artificial. San Francisco, CA: Morgan Kaufmann, 2018: 1043-1049

[26] He Xiangteng, Peng Yuxin. Fine-grained image classification via combining vision and language [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 5994-6002

[27] He Xiangteng, Peng Yuxin. Only learn one sample: Fine-grained visual categorization with one sample training [C] //Proc of the 26th ACM Int Conf on Multimedia. New York: ACM, 2018: 1372-1380

[28] Zhu Chen, Tan Xiao, Zhou Feng, et al. Fine-grained video categorization with redundancy reduction attention [C] //Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2018: 136-152

[29] Gionis A, Indyk P, Motwani R. Similarity search in high dimensions via hashing [C] //Proc of the 25th Int Conf on Very Large Data Bases. New York: ACM, 1999: 518-529

[30] Lü Qin, Josephson W, Wang Zhe, et al. Multi-probe LSH: Efficient indexing for high-dimensional similarity search [C] //Proc of the 33rd Int Conf on Very Large Data Bases. New York: ACM, 2007: 950-961

[31] Satuluri V, Parthasarathy S. Bayesian locality sensitive hashing for fast similarity search [C] //Proc of the 38th Int Conf on Very Large Data Bases. New York: ACM, 2012: 430-441

[32] Liu Wei, Wang Jun, Kumar S, et al. Hashing with graphs [C] //Proc of the 28th Int Conf on Machine Learning. New York: ACM, 2011: 1-8

[33] Weiss Y, Torralba A, Fergus R. Spectral hashing [C] //Proc of the 19th Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2009: 1753-1760

[34] Shen Fumin, Shen Chunhua, Shi Qinfeng, et al. Inductive hashing on manifolds [C] //Proc of the 26th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2013: 1562-1569

[35] Gong Yunchao, Lazebnik S, Gordo A, et al. Iterative quantization: A procrustean approach to learning binary codes [C] //Proc of the 24th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2011: 817-824

[36] Shakhnarovich G, Viola P, Darrell T. Fast pose estimation with parameter-sensitive hashing [C] //Proc of the 2nd IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2003: 750-757

[37] Rastegari M, Farhadi A, Forsyth D. Attribute discovery via predictable discriminative binary code [C] //Proc of the 10th European Conf on Computer Vision. New York: ACM, 2012: 876-889

[38] Liu Wei, Wang Jun, Ji Rongrong, et al. Supervised hashing with kernels [C] //Proc of the 11th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2012: 2074-2081

[39] Mu Yadong, Shen Jialie, Yan Shuicheng. Weakly-supervised hashing in kernel space [C] //Proc of the 9th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2010: 3344-3351

[40] Kulis B, Jain P, Grauman K. Fast similarity search for learned metrics [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(12): 2143-2157

[41] Norouzi M, Blei D. Minimal loss hashing for compact binary codes [C] //Proc of the 28th Int Conf on Machine Learning. New York: ACM, 2011: 353-360

[42] Li Xi, Lin Guosheng, Shen Chunhua, et al. Learning Hash functions using column generation [C] //Proc of the 30th Int Conf on Machine Learning. New York: ACM, 2013: 142-150

[43] Norouzi M, Fleet D, Salakhutdinov R. Hamming distance metric learning [C] //Proc of the 22nd Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2012: 1061-1069

[44] Wang Jianfeng, Wang Jingdong, Yu Nenghai, et al. Order preserving hashing for approximate nearest neighbor search [C] //Proc of the 21st ACM Int Conf on Multimedia. New York: ACM, 2013: 133-142

[45] Wang Jun, Liu Wei, Sun A X, et al. Learning Hash codes with listwise supervision [C] //Proc of the 26th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2013: 3032-3039

[46] Wang Jingdong, Zhang Ting, Song Jingkuan. A survey on learning to Hash [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 769-790

[47] Salakhutdinov R, Hinton G. Semantic hashing [J]. International Journal of Approximate Reasoning, 2009, 50(7): 969-978

[48] Xia Rongkai, Pan Yan, Lai Hanjiang, et al. Supervised hashing for image retrieval via image representation learning [C] //Proc of the 28th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2014: 2156-2162

[49] Lai Hanjiang, Pan Yan, Liu Ye, et al. Simultaneous feature learning and Hash coding with deep neural networks [C] //Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3270-3278

[50] Li Wujun, Wang Sheng, Kang Wangcheng. Feature learning based deep supervised hashing with pairwise labels [C] //Proc of the 25th Int Joint Conf on Artificial. San Francisco, CA: Morgan Kaufmann, 2016: 1711-1717

[51] Liu Haomiao, Wang Ruiping, Shan Shiguang et al. Deep supervised hashing for fast image retrieval [C] //Proc of the 29th IEEE Int Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2064-2072.

[52] Cao Zhangjie, Long Mingsheng, Wang Jianmin et al. HashNet: Deep learning to Hash by continuation [C] //Proc of the 16th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 5609-5618

[53] Zhang Jian, Peng Yuxin. Query-adaptive image retrieval by deep weighted hashing [J]. IEEE Transactions on Multimedia, 2018, 20(9): 2400-2414

[54] Zhang Jian, Peng Yuxin. SSDH: Semi-supervised deep hashing for large scale image retrieval [J/OL]. IEEE Transactions on Circuits and Systems for Video Technology, 2017 [2018-11-01]. https://ieeexplore.ieee.org/document/8101524. DOI: 10.1109/TCSVT.2017.2771332

[55] Laptev I. On space-time interest points [J]. International Journal of Computer Vision, 2005, 64(2/3): 107-123

[56] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C] //Proc of the 18th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2005: 886-893

[57] Laptev I, Marszalek M, Schmid C, et al. Learning realistic human actions from movies [C] //Proc of the 21st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2008: 1-8

[58] Dalal N, Triggs B, Schmid C. Human detection using oriented histograms of flow and appearance [C] //Proc of the 4th European Conf on Computer Vision. Berlin: Springer, 2006: 428-441

[59] Wang Heng, Kläser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition [J]. International Journal of Computer Vision, 2013, 103(1): 60-79

[60] Wang Heng, Schmid C. Action recognition with improved trajectories [C] //Proc of the 12th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2013: 3551-3558

[61] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [C] //Proc of the 24th Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 568-576

[62] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks [C] //Proc of the 15th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 4489-4497

[63] Wang Limin, Xiong Yuanjun, Wang Zhe, et al. Temporal segment networks: Towards good practices for deep action recognition [C] //Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 20-36

[64] Peng Yuxin, Zhao Yunzhen, Zhang Junchao. Two-stream collaborative learning with spatial-temporal attention for video classification [J/OL]. IEEE Transactions on Circuits and Systems for Video Technology. [2018-11-01]. https://ieeexplore. ieee. org/document/8300648. DOI: 10. 1109/TCSVT. 2018. 2808685

[65] Qiu Zhaofan, Yao Ting, Mei Tao. Learning spatio-temporal representation with pseudo-3D residual networks [C] //Proc of the 16th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 5533-5541

[66] Zhao Yue, Xiong Yuanjun, Lin Dahua. Recognize actions by disentangling components of dynamics [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6566-6575

[67] Sun Shuyang, Kuang Zhanghui, Sheng Lu, et al. Optical flow guided feature: A fast and robust motion representation for video action recognition [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 1390-1399

[68] Ng J Y, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification [C] //Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 4694-4702

[69] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 6299-6308

[70] Wang Limin, Li Wei, Li Wen, et al. Appearance-and-relation networks for video classification [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 1430-1439

[71] Wang Xiaolong, Girshick R, Gupta A, et al. Non-local neural networks [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7794-7803

[72] Ma Chihao, Kadav A, Melvin I, et al. Attend and interact: Higher-Order object interactions for video understanding [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6790-6800

[73] Xu Xun, Hospedales T, Gong Shaogang. Transductive zero-shot action recognition by word-vector embedding [J]. International Journal of Computer Vision, 2017, 123(3): 309-333

[74] Gan Chuang, Yang Yi, Zhu Linchao, et al. Recognizing an action using its name: A knowledge-based approach [J]. International Journal of Computer Vision, 2016, 120(1): 61-77

[75] Zhang Chenrui, Peng Yuxin. Visual data synthesis via GAN for zero-shot video classification [C] //Proc of the 27th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2018: 1128-1134

[76] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C] //Proc of the 27th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 580-587

[77] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection [C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 779-788

[78] Kang K, Ouyang W, Li Hognsheng, et al. Object detection from video tubelets with convolutional neural networks [C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 817-825

[79] Zhu Xizhou, Xiong Yuwen, Dai Jifeng, et al. Deep feature flow for video recognition [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 2349-2358

[80] Zhu Xizhou, Wang Yujie, Dai Jifeng, et al. Flow-guided feature aggregation for video object detection [C] //Proc of the 16th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 408-417

[81] Zhu Xizhou, Dai Jifeng, Yuan Lu, et al. Towards high performance video object detection [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7210-7218

[82] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110

[83] Kalantidis Y, Pueyo L G, Trevisiol M, et al. Scalable triangulation-based logo recognition [C] //Proc of the 1st ACM Int Conf on Multimedia Retrieval. New York: ACM, 2011: 20-26

[84] Tang Panpan, Peng Yuxin. Exploiting distinctive topological constraint of local feature matching for logo image recognition [J]. Neurocomputing, 2017, 236: 113-122

[85] Yang Yi, Luo Hengliang, Xu Huarong, et al. Towards real-time traffic sign detection and classification [J]. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(7): 2022-2031

[86] Zhu Chao, Peng Yuxin. Discriminative latent semantic feature learning for pedestrian detection [J]. Neurocomputing, 2017, 238: 126-138

[87] Zhu Chao, Peng Yuxin. A boosted multi-task model for pedestrian detection with occlusion handling [C] //Proc of the 29th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2015: 3878-3884

[88] Zhu Chao, Peng Yuxin. A boosted multi-task model for pedestrian detection with occlusion handling [J]. IEEE Transactions on image processing, 2015, 24(12): 5619-5629

[89] Zhu Chao, Peng Yuxin. Group cost-sensitive boosting for multi-resolution pedestrian detection [C] //Proc of the 30th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2016: 3676-3682

[90] Li Jianan, Liang Xiaodan, Shen Sengmei, et al. Scale-aware fast R-CNN for pedestrian detection [J]. IEEE Transactions on Multimedia, 2018, 20(4): 985-996

[91] Zhang Shanshan, Yang Jian, Schiele B. Occluded pedestrian detection through guided attention in CNNs [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6995-7003

[92] Yi Jian, Peng Yuxin, Xiao Jianguo. Color-based clustering for text detection and extraction in image [C] //Proc of the 15th ACM Int Conf on Multimedia. New York: ACM, 2007: 847-850

[93] Mishra A, Alahari K, Jawahar C V. Top-down and bottom-up cues for scene text recognition [C] //Proc of the 25th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2012: 2687-2694

[94] Shi Baoguang, Wang Xinggang, Lyu Pengyuan, et al. Robust scene text recognition with automatic rectification [C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 4168-4176

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

[96] Hotelling H. Relations between two sets of variates [J]. Biometrika, 1936, 28(3/4): 321-377

[97] Rasiwasia N, Mahajan D, Mahadevan V, et al. Cluster canonical correlation analysis [C] //Proc of the 17th Int Conf on Artificial Intelligence and Statistics. Reykjavik: JMLR, 2014: 823-831

[98] Gong Yunchao, Ke Qifa, Isard M, et al. A multi-view embedding space for modeling Internet images, tags, and their semantics [J]. International Journal of Computer Vision, 2014, 106(2): 210-233

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

[100] Li Dongge, Dimitrova N, Li Mingkun, et al. Multimedia content processing through cross-modal association [C] //Proc of the 11th ACM Int Conf on Multimedia. New York: ACM, 2003: 604-611

[101] Zhai Xiaohua, Peng Yuxin, Xiao Jiangguo. Heterogeneous metric learning with joint graph regularization for cross-media retrieval [C] //Proc of the 27th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2013: 1198-1204

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

[103] Peng Yuxin, Zhai Xiaohua, Zhao Yunzhen, et al. Semi-supervised cross-media feature learning with unified patch graph regularization [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 26(3): 583-596

[104] Grangier D, Bengio S. A discriminative kernel-based model to rank images from text queries [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(8): 1371-1384

[105] Wu Fei, Lu Xinyan, Zhang Zhongfei, et al. Cross-media semantic representation via bi-directional learning to rank [C] //Proc of the 21st ACM Int Conf on Multimedia. New York: ACM, 2013: 877-886

[106] Zhuang Yueting, Wang Yanfei, Wu Fei, et al. Supervised coupled dictionary learning with group structures for multi-modal retrieval [C] //Proc of the 27th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2013: 1070-1076

[107] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning [C] //Proc of the 28th Int Conf on Machine Learning. New York: ACM, 2011: 689-696

[108] Srivastava N, Salakhutdinov R R. Multimodal learning with deep Boltzmann machines [C] //Proc of the 22nd Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2012: 2222-2230

[109] Andrew G, Arora R, Bilmes J, et al. Deep canonical correlation analysis [C] //Proc of the 30th Int Conf on Machine Learning. New York: ACM, 2013: 1247-1255

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

[111] Wang Wei, Ooi B C, Yang Xiaoyan, et al. Effective multi-modal retrieval based on stacked auto-encoders [J]. Proceedings of the VLDB Endowment, 2014, 7(8): 649-660

[112] Peng Yuxin, Huang Xin, Qi Jinwei. Cross-media shared representation by hierarchical learning with multiple deep networks [C] //Proc of Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2016: 3846-3853

[113] Peng Yuxin, Qi Jinwei, Huang Xin, et al. CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network [J]. IEEE Transactions on Multimedia, 2018, 20(2): 405-420

[114] Wei Yunchao, Zhao Yao, Lu Canyi, et al. Cross-modal retrieval with CNN visual features: A new baseline [J]. IEEE Transactions on Cybernetics, 2017, 47(2): 449-460

[115] Peng Yuxin, Qi Jinwei, Yuan Yuxin. Modality-specific cross-modal similarity measurement with recurrent attention network [J]. IEEE Transactions on Image Processing, 2018, 27(11): 5585-5599

[116] Qi Jinwei, Peng Yuxin. Cross-modal bidirectional translation via reinforcement learning [C] //Proc of the 27th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2018: 2630-2636

[117] Qi Jinwei, Peng Yuxin, Yuan Yuxin. Cross-media multi-level alignment with relation attention network [C] //Proc of the 27th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2018: 892-898

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

[119] Gu Jiuxiang, Cai Jianfei, Joty S, et al. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7181-7189

[120] Fan Mengdi, Wang Wenmin, Dong Peilei, et al. Cross-media retrieval by learning rich semantic embeddings of multimedia [C] //Proc of the 25th ACM Int Conf on Multimedia. New York: ACM, 2017: 1698-1706

[121] Huang Xin, Peng Yuxin, Yuan Mingkuan. Cross-modal common representation learning by hybrid transfer network [C] //Proc of the 26th Int Joint Conf Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2017: 1893-1900

[122] Huang Xin, Peng Yuxin. Deep cross-media knowledge transfer [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8837-8846

[123] Song Jingkuan, Yang Yang, Yang Yi, et al. Inter-media hashing for large-scale retrieval from heterogeneous data sources [C] //Proc of the 40th ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2013: 785-796

[124] Long Mingsheng, Cao Yue, Wang Jianmin, et al. Composite correlation quantization for efficient multimodal retrieval [C] //Proc of the 39th Int ACM SIGIR Conf on Research and Development in Information. New York: ACM, 2016: 579-588

[125] Liu Hong, Ji Rongrong, Wu Yongjian, et al. Cross-modality binary code learning via fusion similarity hashing [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 6345-6353

[126] Bronstein M M, Bronstein A M, Michel F, et al. Data fusion through cross-modality metric learning using similarity-sensitive hashing [C] //Proc of the 23rd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2010: 3594-3601

[127] Wei Ying, Song Yangqiu, Zhen Yi, et al. Scalable heterogeneous translated hashing [C] //Proc of the 20th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 791-800

[128] Lin Zijia, Ding Guiguang, Hu Mingqing, et al. Semantics-preserving hashing for cross-view retrieval [C] //Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3864-3872

[129] Zhuang Yueting, Yu Zhou, Wang Wei, et al. Cross-media hashing with neural networks [C] //Proc of the 22nd ACM Int Conf on Multimedia. New York: ACM, 2014: 901-904

[130] Ye Zhaoda, Peng Yuxin. Multi-Scale correlation for sequential cross-modal hashing learning [C] //Proc of the 25th ACM Int Conf on Multimedia. New York: ACM, 2018: 852-860

[131] Zhang Jian, Peng Yuxin, Yuan Mingkuan, SCH-GAN: Semi-supervised cross-modal hashing by generative adversarial network [J]. IEEE Transactions on Cybernetics, 2018

[132] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C] //Proc of the 24th Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 2672-2680

[133] Zhang Jian, Peng Yuxin, Yuan Mingkuan. Unsupervised generative adversarial cross-modal hashing [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 539-546

[134] Kulkarni G, Premraj V, Ordonez V, et al. Babytalk: Understanding and generating simple image descriptions [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2891-2903

[135] Yang Yezhou, Teo C L, Daumé III H, et al. Corpus-guided sentence generation of natural images [C] //Proc of the 8th Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2011: 444-454

[136] Ordonez V, Kulkarni G, Berg T L. Im2text: Describing images using 1 million captioned photographs [C] //Proc of the 25th Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2011: 1143-1151

[137] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator [C] //Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3156-3164

[138] Venugopalan S, Xu Huijuan, Donahue J, et al. Translating videos to natural language using deep recurrent neural networks [C] //Proc of the 2015 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2014: 1494-1504

[139] Venugopalan S, Rohrbach M, Donahue J, et al. Sequence to sequence-video to text [C] //Proc of the 15th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 4534-4542

[140] Xu K, Ba J L, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention [C] //Proc of the 32nd Int Conf on Machine Learning. New York: ACM, 2015: 2048-2057

[141] Zhang Junchao, Peng Yuxin. Hierarchical vision-language alignment for video captioning [C/OL] //Proc of the 25th Int Conf on MultiMedia Modeling. [2018-11-15]. http://mmm2019.iti.gr/conference-program/

[142] Hori C, Hori T, Lee T Y, et al. Attention-based multimodal fusion for video description [C] //Proc of the 16th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 4203-4212

[143] Venugopalan S, Hendricks L A, Rohrbach M, et al. Captioning images with diverse objects [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 5753-5761

[144] Aneja J, Deshpande A, Schwing A G. Convolutional image captioning [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 5561-5570

[145] Ranzato M, Chopra S, Auli M, et al. Sequence level training with recurrent neural networks [OL]. [2018-07-30]. https://arxiv.org/pdf/1511.06732.pdf

[146] Liu Siqi, Zhu Zhenhai, Ye Ning, et al. Improved image captioning via policy gradient optimization of SPIDEr [C] //Proc of the 16th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 873-881

[147] Pasunuru R, Bansal M. Reinforced video captioning with entailment rewards [C] //Proc of the 14th Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2017: 979-985

[148] Johnson J, Karpathy A, Li Feifei. DenseCap: Fully convolutional localization networks for dense captioning [C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 4565-4574

[149] Krishna R, Hata K, Ren F, et al. Dense-captioning events in videos [C] //Proc of the 16th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 706-715

[150] Yu Huanyu, Cheng Shuo, Ni Bingbing, et al. Fine-grained video captioning for sports narrative [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6006-6015

[151] Lan Weiyu, Li Xirong, Dong Jianfeng. Fluency-guided cross-lingual image captioning [C] //Proc of the 25th ACM Int Conf on Multimedia. New York: ACM, 2017: 1549-1557

[152] Kingma D P, Welling M. Auto-encoding variational Bayes [OL]. [2018-07-20]. http://arxiv.org/abs/1312.6114

[153] Yan Xinchen, Yang Jimei, Sohn K, et al. Attribute2image: Conditional image generation from visual attributes [C] //Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 776-791

[154] Reed S, Akata Z, Yan Xinchen, et al. Generative adversarial text to image synthesis [C] //Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016: 1060-1069

[155] Reed S E, Akata Z, Mohan S, et al. Learning what and where to draw [C] //Proc of the 30th Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2016: 217-225

[156] Zhang Han, Xu Tao, Li Hongsheng. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks [C] //Proc of the 16th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 5907-5915

[157] Huang Xun, Li Yixuan, Poursaeed O, et al. Stacked generative adversarial networks [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 5077-5086

[158] Zhang Zizhao, Xie Yuanqu, Yang Lin. Photographic text-to-image synthesis with a hierarchically-nested adversarial network [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6199-6208

[159] Xu Tao, Zhang Pengchuan, Huang Qiuyuan, et al. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 1316-1324

[160] Yuan Mingkuan, Peng Yuxin. Text-to-image synthesis via symmetrical distillation networks [C] //Proc of the 26th ACM Int Conf on Multimedia. New York: ACM, 2018: 1407-1415

[161] Johnson J, Gupta A, Li Feifei. Image generation from scene graphs [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 1219-1228

[162] Mittal G, Marwah T, Balasubramanian V N. Sync-DRAW: Automatic video generation using deep recurrent attentive architectures [C] //Proc of the 25th ACM Int Conf on Multimedia. New York: ACM, 2017: 1096-1104

[163] Marwah T, Mittal G, Balasubramanian V N. Attentive semantic video generation using captions [C] //Proc of the 16th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 1462-1434

[164] Pan Yingwei, Qiu Zhaofan, Yao Ting, et al. To create what you tell: Generating videos from captions [C] //Proc of the 25th ACM Int Conf on Multimedia. New York: ACM, 2017: 1789-1798

[165] Li Yitong, Min M R, Shen Dinghan, et al. Video generation from text [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 7065-7072

[166] Antol S, Agrawal A, Lu Jiasen, et al. VQA: Visual question answering [C] //Proc of the 15th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 2425-2433

[167] Xue Hongyang, Zhao Zhou, Cai Deng. Unifying the video and question attentions for open-ended video question answering [J]. IEEE Transactions on Image Processing, 2017, 26(12): 5656-5666

[168] Geman D, Geman S, Hallonquist N, et al. Visual turing test for computer vision systems [J]. Proceedings of the National Academy of Sciences, 2015, 112(12): 3618-3623

[169] Wu Qi, Teney D, Wang Peng, et al. Visual question answering: A survey of methods and datasets [J]. Computer Vision and Image Understanding, 2017, 163: 21-40

[170] Kafle K, Kanan C. Visual question answering: Datasets, algorithms, and future challenges [J]. Computer Vision and Image Understanding, 2017, 163: 3-20

[171] Lu Jiasheng, Yang Jianwei, Batra D, et al. Hierarchical question-image co-attention for visual question answering [C] //Proc of the 26th Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2016: 289-297

[172] Andreas J, Rohrbach M. Darrell T, et al. Deep compositional question answering with neural module networks [OL]. [2018-07-24]. https://arxiv. org/pdf/1511.02799. pdf

[173] Santoro A, Raposo D, Barrett D G T, et al. A simple neural network module for relational reasoning [C] //Proc of the 27th Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2017: 4974-4983

[174] Wu Qi, Wang Peng, Shen Chunhua, et al. Ask me anything: Free-form visual question answering based on knowledge from external sources [C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 4622-4630

[175] Su Zhou, Zhu Chen, Dong Yipeng, et al. Learning visual knowledge memory networks for visual question answering [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7736-7745

[176] Wang Peng, Wu Qi, Shen Chunhua, et al. Explicit knowledge-based reasoning for visual question answering [C] //Proc of the 26th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2017: 1290-1296

[177] Aditya S, Yang Yezhou, Baral C. Explicit reasoning over end-to-end neural architectures for visual question answering [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 629-637

Peng Yuxin, born in 1974. PhD, professor. Senior member of CCF. His main research interests include cross-media analysis and reasoning, image and video analysis and retrieval, and computer vision.

Qi Jinwei, born in 1994. Master candidate. His main research interests include cross-media retrieval and machine learning.

Huang Xin, born in 1991. PhD candidate. His main research interests include cross-media reasoning and machine learning.