·热点与综述· 文章编号:1000-3428(2020)05-0001-11 文献标志码:A 中图分类号:TP391.1

面向深度学习的多模态融合技术研究综述

何 俊[1],张彩庆[2a],李小珍[1],张德海[2b]

(1. 昆明学院 信息工程学院,昆明 650214; 2. 云南大学 a. 外国语学院; b. 软件学院,昆明 650206)

摘 要: 面向深度学习的多模态融合技术是指机器从文本、图像、语音和视频等领域获取信息实现转换与融合以提升模型性能,而模态的普遍性和深度学习的热度促进了多模态融合技术的发展。在多模态融合技术发展前期,以提升深度学习模型分类与回归性能为出发点,阐述多模态融合架构、融合方法和对齐技术。重点分析联合、协同、编解码器 3 种融合架构在深度学习中的应用情况与优缺点,以及多核学习、图像模型和神经网络等具体融合方法与对齐技术,在此基础上归纳多模态融合研究的常用公开数据集,并对跨模态转移学习、模态语义冲突消解、多模态组合评价等下一步的研究方向进行展望。

关键词: 深度学习;多模态;模态融合;模态对齐;多核学习;图像模型

开放科学(资源服务)标志码(OSID):

中文引用格式: 何俊,张彩庆,李小珍,等. 面向深度学习的多模态融合技术研究综述[J]. 计算机工程,2020,46(5):1-11.

英文引用格式: HE Jun, ZHANG Caiqing, LI Xiaozhen, et al. Survey of research on multimodal fusion technology for deep learning[J]. Computer Engineering, 2020, 46(5):1-11.

Survey of Research on Multimodal Fusion Technology for Deep Learning

HE Jun[1], ZHANG Caiqing[2a], LI Xiaozhen[1], ZHANG Dehai[2b]

(1. College of Information Engineering, Kunming University, Kunming 650214, China;
2a. College of Foreign Languages; 2b. College of Software, Yunnan University, Kunming 650206, China)

【Abstract】 Multimodal Fusion Technology (MFT) for Deep Learning (DL) refers to the conversion and fusion of information obtained by machine from texts, images, voices, videos and other materials, so as to improve the performance of the model. The universality of modals and the heat of DL boost the rapid development of multimodal fusion. In order to improve the performance of DL model classification or regression, this paper summarizes the multimodal fusion architecture, fusion methods and alignment technologies in the early stage of MFT development. This paper focuses on the analysis of the three fusion architectures: joint, cooperative and codec architectures, in terms of their adoption in DL and advantages/disadvantages. The specific fusion methods and alignment technologies such as Multiple Kernel Learning (MKL), Graphic Model (GM) and Neural Network (NN) are also studied. Finally, the public datasets commonly used in multimodal fusion research are summarized, and the direction of further research in cross-modal transfer learning, resolution of modal semantic conflicts, and multimodal combination evaluation is prospected.

【Key words】 Deep Learning (DL); multimodality; modal fusion; modal alignment; Multiple Kernel Learning (MKL); Graphical Model (GM)

DOI: 10.19678/j.issn.1000-3428.0057370

0 概述

近年来,深度学习(Deep Learning, DL)在图像识别、机器翻译、情感分析、自然语言处理(Natural Language Processing, NLP)等领域得到广泛应用并取得较多研究成果,为使机器能更全面高效地感知周围的世界,需要赋予其理解、推理及融合多模态信息的能力,并且由于人们生活在一个多领域相互交

基金项目: 国家自然科学基金(61263043,61864004);云南省地方本科高校基础研究联合专项(2017FH001-05)。

作者简介: 何 俊(1977—),男,副教授、博士,主研方向为机器学习、软件演化、数据分析;张彩庆,讲师、硕士;李小珍,讲师、博士;张德海,副教授、博士。

收稿日期: 2020-02-11 **修回日期:** 2020-03-13 **E-mail:** 369885901@qq.com

融的环境中,听到的声音、看到的实物、闻到的味道都是一种模态,因此研究人员开始关注如何将多领域数据进行融合实现异质互补,例如语音识别的研究表明,视觉模态提供了嘴的唇部运动和发音信息,包括张开和关闭,有助于提高语音识别性能。可见,利用多种模式的综合语义对深度学习研究具有重要意义。深度学习中的多模态融合技术(Multimodality Fusion Technology,MFT)[1]是模型在分析和识别任务时处理不同形式数据的过程。多模态数据的融合可为模型决策提供更多信息,从而提高决策总体结果的准确率,其目标是建立能够处理和关联来自多种模态信息的模型。

MFT 主要包括模态表示、融合、转换、对齐技术[2]。由于不同模态的特征向量最初位于不同的子空间中,即具有异质性,因此将影响多模态数据在深度学习领域的应用[3]。为解决该问题,可将异构特征投影到公共子空间,由相似向量表示具有相似语义的多模态数据[4]。因此,多模态融合技术的主要目标是缩小语义子空间中的分布差距,同时保持模态特定语义的完整性,例如利用多模态融合特征,提高视频分类[5]、事件检测[6-7]、情感分析[8-9]、跨模态翻译[10]等跨媒体分析性能。特别是多模态融合近期在计算机视觉、NLP 和语音识别等应用中取得的突出性成果[11],已引起学术界和工业界的广泛关注。本文根据多模态融合架构、融合方法、模态对齐方式和公开数据资源等,对面向深度学习的多模态融合技术进行分析与研究。

1 多模态融合架构

多模态融合的主要目标是缩小模态间的异质性差异,同时保持各模态特定语义的完整性,并在深度学习模型中取得较优的性能。多模态融合架构分为[2]:联合架构,协同架构和编解码器架构。联合架构是将单模态表示投影到一个共享语义子空间中,以便能够融合多模态特征。协同架构包括跨模态相似模型和典型相关分析,其目标是寻找协调子空间中模态间的关联关系。编解码器架构是将一个模态映射到另一个模态的多模态转换任务中。3 种融合架构在视频分类、情感分析、语音识别等领域得到广泛应用,且涉及图像、视频、语音、文本等融合内容,具体应用情况如表 1 所示。

表 1 3 种多模态融合架构的应用情况

Table 1 Application situation of three architectures for multimodal fusion

架构	应用领域	融合内容	参考文献
联合架构	视频分类	语音、视频、文本	文献[5,12]
	事件检测	语音、视频、文本	文献[7]
	情绪分析	语音、视频、文本	文献[13-14]
	视觉问答	图像、文本	文献[15-16]
	情感分析	语音、视频、文本	文献[17]
	语音识别	语音、视频	文献[18]
协同架构	跨模态搜索	图像、文本	文献[19-20]
	图像标注	图像、文本	文献[21]
	跨模态嵌入	图像、视频、文本	文献[22-23]
	转移学习	图像、文本	文献[24]
编解码器架构	图像标注	图像、文本	文献[25]
	视频解码	视频、文本	文献[26-27]
	图像合成	图像、文本	文献[28]

1.1 联合架构

多模态融合策略是集成不同类型的特征来提高机器学习模型性能,消除不同模态的异质性差异。联合架构是将多模态空间映射到共享语义子空间中,从而融合多个模态特征[2],如图 1 所示。每个单一模态通过单独编码后,将被映射到共享子空间中,遵循该策略,其在视频分类[12]、事件检测[7]、情感分析[13-14]、视觉问答[15-16]和语音识别[17-18]等多模态分类或回归任务中都表现出较优的性能。

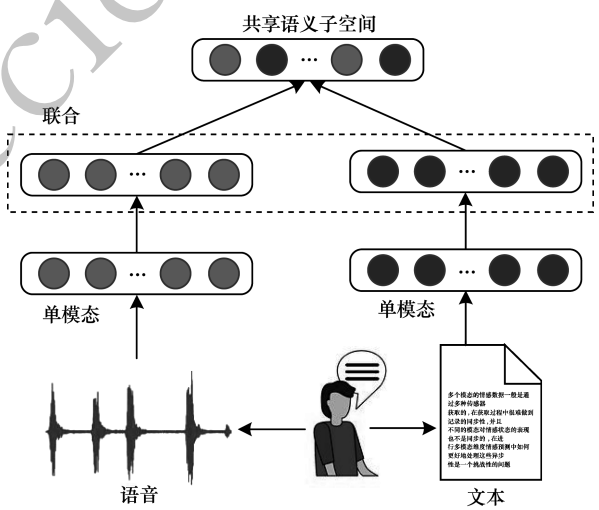


图 1 联合融合架构示意图

Fig. 1 Schematic diagram of joint fusion architecture

多模态联合架构的关键是实现特征“联合”,一种较简单的方法是直接连接,即“加”联合方法。该方法在不同的隐藏层实现共享语义子空间,将转换后的各个单模态特征向量语义组合在一起,从而实

现多模态融合,如式(1)所示:

$$z = f(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{v}_1 + \boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{v}_2 + \cdots + \boldsymbol{w}_n^{\mathrm{T}}\boldsymbol{v}_n) \tag{1}$$

其中,z 是共享语义子空间中的输出结果,$\boldsymbol{v}$ 是各单模态的输入,$\boldsymbol{w}$ 是权重,下标表示不同的模态,通过映射 f 将所有子模态语义转换到共享子空间。

另一种常用方法是“乘”联合方法,如文献[29]将语言、视频和音频等模态融合在统一的张量中,而张量是由所有单模态特征向量的输出乘积构成,如式(2)所示:

$$z = \begin{bmatrix} \boldsymbol{v}^1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{v}^2 \\ 1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} \boldsymbol{v}^n \\ 1 \end{bmatrix} \tag{2}$$

其中,z 表示融合张量后的结果输出,$\boldsymbol{v}$ 表示不同的模态,$\otimes$表示外积算子。

尽管“加”联合方法简单且容易实现,但其特征向量语义组合容易造成后期语义丢失,使模型性能降低,而“乘”联合方法弥补了这一不足,通过张量计算使特征语义得到充分融合,例如文献[17]的多模态情感预测模型由包括许多内部乘积的连续神经层组成,其充分利用深度神经网络的多层性质,将不同模态有序分布在不同层中,并在模型训练过程中动态实现向量语义组合。

此外,联合架构对每个单模态的语义完整性有较高要求,数据不完整或错误问题在后期融合中会被放大,一些研究人员通过联合训练或模态相关性来解决这一问题。文献[30-31]通过多模态联合处理某些单模态中的部分数据缺失问题,以便可以利用更多且更完整的训练数据,或者在一种或多种模态数据缺失的情况下,尽量减少对后续训练任务的影响。文献[12]利用各单模态特征之间的相关性(如权重相似性)来发现模态之间的关系,从而对这些特征进行分类使用,该方法在视频分类任务中的实验结果表明其有助于提高机器学习模型性能。

多模态联合架构的优点是融合方式简单,且共享子空间通常具备语义不变性,有助于在机器学习模型中将知识从一种模态转换到另一种模态。其缺点是各单模态语义完整性不易在早期发现和处理。

1.2　协同架构

多模态协同架构是将各种单模态在一些约束的作用下实现相互协同[2]。由于不同模态包含的信息不同,因此协同架构有利于保持各单模态独有的特征和排它性,如图 2 所示。

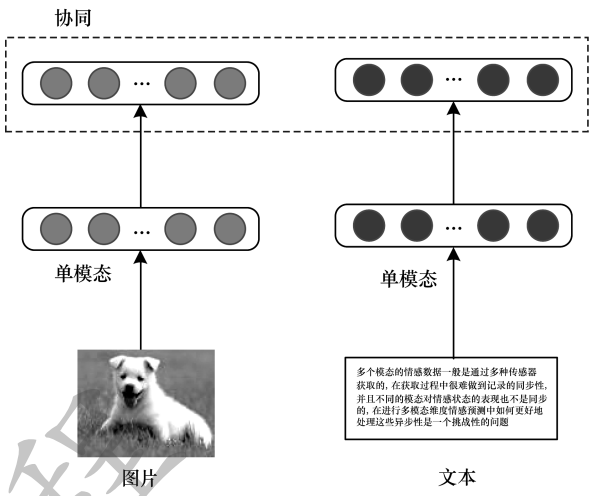


图 2　协同融合架构示意图

Fig. 2　Schematic diagram of collaborative fusion architecture

协同架构在跨模态学习中已经得到广泛应用,主流的协同方法是基于交叉模态相似性方法,该方法旨在通过直接测量向量与不同模态的距离来学习公共子空间[32]。基于交叉模态相关性的方法旨在学习一个共享子空间,从而使不同模态表示集的相关性最大化[4]。

交叉模态相似性方法在相似性度量的约束下保持模态间和模态内的相似性结构,使得相同语义或相关对象的跨模态相似距离尽可能小,不同语义的距离尽可能大,例如文献[23]提出的模态间排名方法用于完成视觉和文本融合任务,将视觉和文本的匹配嵌入向量表示为$(\boldsymbol{v},\boldsymbol{t}) \in D$,融合目标函数用一个损失函数 f 表示,如式(3)所示:

$$f = \sum_{\boldsymbol{v}} \sum_{\boldsymbol{t}^-} \max(0, \alpha - S(\boldsymbol{v},\boldsymbol{t}) + S(\boldsymbol{v},\boldsymbol{t}^-)) + \sum_{\boldsymbol{t}} \sum_{\boldsymbol{v}^-} \max(0, \alpha - S(\boldsymbol{t},\boldsymbol{v}) + S(\boldsymbol{t},\boldsymbol{v}^-)) \tag{3}$$

其中,α 是边缘,S 是相似性度量函数,$\boldsymbol{t}^-$ 是与 $\boldsymbol{v}$ 不匹配的嵌入向量,$\boldsymbol{v}^-$ 是与 $\boldsymbol{t}$ 不匹配的嵌入向量,且 $\boldsymbol{t}^-$ 和 $\boldsymbol{v}^-$ 是随机选择的样本。该方法保持了模态间和模态内的相似性结构,同时实现模态之间相互协同。此外,文献[22,33-34]采用其他方法来度量距离,如欧式距离,其目的都是使配对样本距离最小化。除了学习模态间相似性的度量外,跨模态应用的另一个关键问题是保持模态间相似性结构,此类方法通常对模态特征的类别进行分类,使它们在每种模态下具有一定的区分度[19],同时兼顾模态协同和特征融合。由于协同架构的这一灵活特点,使其在语音识别、迁移学习和图像标注等领域都有广泛应用。

协同架构的优点是每个单模态都可以独立运行,这一特性有利于跨模式迁移学习,其目的是在不

同模态或领域之间传递知识。其缺点是模态融合难度较大,使跨模态学习模型不容易实现,同时模型很难在两种以上的模态之间实现迁移学习。

1.3 编解码器架构

编解码器架构通常用于将一种模态映射到另一种模态的多模态转换任务中,主要由编码器和解码器两部分组成。编码器将源模态映射到向量 $\boldsymbol{v}$ 中,解码器基于向量 $\boldsymbol{v}$ 生成一个新的目标模态样本。该架构在图像标注、图像合成、视频解码等领域有广泛应用,如图 3 所示。

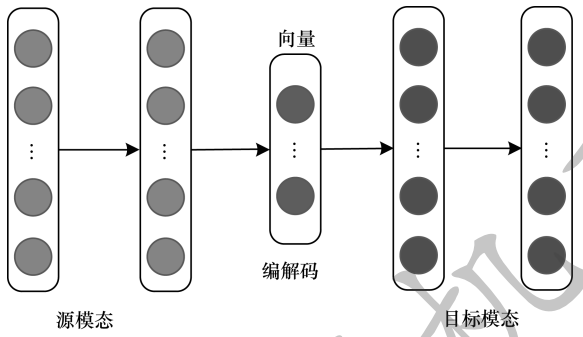


图 3 编解码器融合架构示意图

Fig.3 Schematic diagram of encoder-decoder fusion architecture

目前,编解码器架构重点关注共享语义捕获和多模序列的编解码问题。为有效捕获源模态和目标模态两种模态的共享语义,主流的解决方案是通过一些正则化术语保持模态之间的语义一致性,需确保编码器能正确检测和编码信息,而解码器能推理高级语义和生成语法,以保证源模态中语义的正确理解和目标模态中新样本的生成。为解决多模序列的编码和解码问题,需训练一个灵活的特征选择模块,而训练序列的编码或解码可以看作顺序决策问题,因此通常需采用决策能力强的模型和方法处理该问题,例如深度强化学习(Deep Reinforcement Learning,DRL),其是一种常用的多模序列编解码工具[35]。

尽管多数编解码器架构只包含编码器和解码器,但也有一些架构是由多个编码器或解码器组成。例如:文献[36]提出一种跨乐器翻译音乐的模型,其中涉及一个编码器和多个解码器;文献[37]是一种图像到图像的翻译模型,由多个内容编码器和样式编码器组成,每个编码器都负责一部分工作。

编解码器架构的优点是能够在源模态基础上生成新的目标模态样本。其缺点是每个编码器和解码器只能编码其中一种模态,并且决策模块设计复杂。

2 多模态融合方法

多模态融合方法是多模态深度学习技术的核心内容,本文将从融合技术的角度出发对早期、晚期和混合融合方法[38-39]进行分析。多模态融合方法如表 2所示。

表 2 多模态融合方法

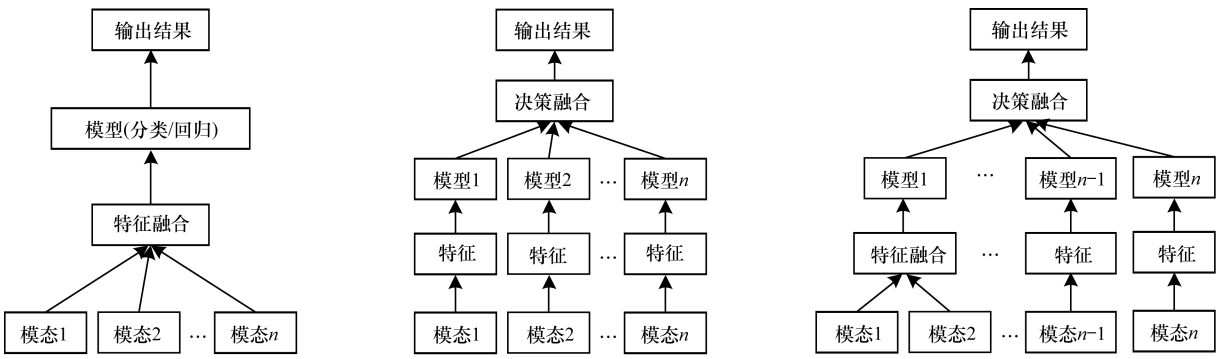
Table 2 Multimodal fusion methods

融合方法	融合类型	输出	时序模型	典型应用	参考文献
模型无关的方法	早期融合	分类	否	情感识别	文献[40]
	晚期融合	回归	是	情感识别	文献[41]
	混合融合	分类	否	事件检测	文献[42]
基于模型的方法	多核学习	分类	否	对象分类	文献[43]
		分类	否	情感识别	文献[44-45]
	图像模型	分类	是	双模语音	文献[46]
		回归	是	情感识别	文献[47]
		分类	否	媒体分类	文献[48]
	神经网络	分类	是	情感识别	文献[49-50]
		分类	否	双模语音	文献[30]
		回归	是	情感识别	文献[51]

将多模态融合方法分为模型无关的方法和基于模型的方法,前者不直接依赖于特定的深度学习方法,后者利用深度学习模型显式地解决多模态融合问题,例如多核学习(Multiple Kernel Learning,MKL)方法、图像模型(Graphical Model,GM)方法和神经网络(Neural Network,NN)方法等。

2.1 模型无关的融合方法

模型无关的融合方法可以分为早期融合(基于特征)、晚期融合(基于决策)和混合融合[11]。如图 4所示,早期融合在提取特征后立即集成特征(通常只需连接各模态特征的表示),晚期融合在每种模式输出结果(例如输出分类或回归结果)后才执行集成,混合融合结合早期融合方法和单模态预测器的输出。



(a) 早期融合方法　　(b)晚期融合方法　　(c)混合融合方法

图 4　3 种模型无关的多模态融合方法

Fig. 4　Three model-independent multimodal fusion methods

2.1.1　早期融合方法

为缓解各模态中原始数据间的不一致性问题,可以先从每种模态中分别提取特征的表示,然后在特征级别进行融合,即特征融合。由于深度学习本质上会涉及从原始数据中学习特征的具体表示,从而导致有时需在未抽取特征之前就进行数据融合,因此特征层面和数据层面的融合均称为早期融合。

模态之间通常是高度相关的,但这种相关性在特征层和数据层提取难度很大。文献[52]认为,不同的数据流所包含的信息之间在较高层次才能具有相关性。文献[53]提出多模态数据的早期融合不能充分展示模态之间的互补性,但可能导致冗余向量的输入。因此,研究人员通常采用降维技术来消除输入空间中的冗余问题,例如文献[54]中的主成分分析(Principal Component Analysis,PCA)方法被广泛应用于多模态深度学习的降维处理中。此外,多模态早期融合方法还需解决不同数据源之间的时间同步问题,文献[55]提出多种解决同步问题的方法,如卷积、训练和池融合等,能较好地将离散事件序列与连续信号进行整合,实现模态间的时间同步。

2.1.2　晚期融合方法

晚期融合方法也称为决策级融合方法,深度学习模型先对不同模态进行训练,再融合多个模型输出的结果。因为该方法的融合过程与特征无关,且来自多个模型的错误通常是不相关的,因此该融合方法普遍受到关注。目前,晚期融合方法主要采用规则来确定不同模型输出结果的组合,即规则融合,例如最大值融合、平均值融合、贝叶斯规则融合以及集成学习等规则融合方法[56]。文献[55]尝试将早期和晚期融合方法进行比较,发现当模态之间相关性比较大时晚期融合优于早期融合,当各个模态在很大程度上不相关时,例如维数和采样率极不相关,采用晚期融合方法则更适合。因此,两种方法各有优缺点,需要在实际应用中根据需求选择。

2.1.3　混合融合方法

混合融合方法结合了早期和晚期融合方法,在综合两者优点的同时,也增加了模型的结构复杂度和训练难度。由于深度学习模型结构的多样性和灵活性,比较适合使用混合融合方法,因此在多媒体、视觉问答、手势识别[57]等领域应用广泛。文献[58]在视频和声音信号融合过程中,先进行仅基于视频信号和声音信号的视听深度神经网络模型训练,分别产生模型预测结果,再将视频信号和声音信号的集成特征输入视听深度神经网络模型中产生模型预测结果,最后采用加权方式整合各模型的预测结果,获得最终识别结果。混合融合方法的组合策略的合理性问题是提高模型性能的关键因素。文献[42]利用混合融合方法实现多媒体事件检测的典型应用,通过早期融合与晚期融合来捕捉特征关系和处理过拟合问题,设计双融合的混合融合方案,达到 88.1% 的准确率,是目前该领域取得的最优结果。

综上,3 种融合方法各有优缺点,早期融合能较好地捕捉特征之间的关系,但容易过度拟合训练数据。晚期融合能较好地处理过拟合问题,但不允许分类器同时训练所有数据。尽管混合多模态融合方法使用灵活,但研究人员针对当前多数的体系结构需根据具体应用问题和研究内容选择合适的融合方法。

2.2　基于模型的融合方法

基于模型的融合方法是从实现技术和模型的角度解决多模态融合问题,常用方法包括 MKL、GM、NN 方法等。

2.2.1　多核学习方法

MKL 是内核支持向量机(Support Vector Machine,SVM)方法的扩展,其允许使用不同的核对应数据的不同视图[59]。由于核可以看作各数据点之间的相似函数,因此该方法能更好地融合异构数据且使用灵活,在多目标检测[43]、多模态情感识别[44]和多模态情感分析[45]等领域均具有非常广泛的应用。文献[60]使用 MKL 从声学、语义和社会学等数据中进行音乐艺术家相似性排序,将异构数据集成到一

个单一、统一的相似空间中,该方法较符合人类的感知。文献[61]在阿尔茨海默病分类中使用 MKL 进行多模态融合,通过在高斯核上进行傅里叶变换,显式计算映射函数,从而得到一个更简单的解决方案,其是一种较新的多核学习框架。这两个研究成果都具有可扩展性和易于实现的特点,并取得了非常出色的学习性能。

除了核选择的灵活性外,MKL 的另一个优势是损失函数为凸,允许使用标准优化包和全局最优解进行模型训练,可大幅提升深度神经网络模型性能。MKL 的主要缺点是在测试期间需要依赖训练数据,且占用大量内存资源。

2.2.2 图像模型方法

GM 是一种常用的多模态融合方法,主要通过图像分割、拼接和预测对浅层或深度图形进行融合,从而生成模态融合结果。常见图像模型有联合概率生成模型和条件概率判别模型[62]等。早期人们多数使用生成模型进行多模态融合,如耦合和阶乘隐马尔可夫模型、动态贝叶斯网络等,这些模型充分利用联合概率的预测能力进行建模,但不利于实现数据的空间和时间结构。近期提出的条件随机场(Conditional Random Fields,CRF)方法通过结合图像描述的视觉和文本信息,可以更好地分割图像[63],并在多模态会议分割[64]、多视点隐藏[65]、潜在变量模型[66]、多媒体分类任务、连续版本的数据拟合等方面都有较好的融合效果。GM 方法利用回归模型对多个连续版本的数据进行拟合,预测后续版本数据的趋势,从而提高多媒体分类任务的性能。

GM 融合方法的优点是能够有效利用数据空间和时间结构,适用于与时间相关的建模任务,还可将人类专家知识嵌入到模型中,增强了模型的可解释性,但是模型的泛化能力有限。

2.2.3 神经网络方法

NN 是目前应用最广泛的方法之一,已用于各种多模态融合任务中[30]。视觉和听觉双模语音识别(Audio-Visual Speech Recognition,AVSR)是最早使用神经网络方法进行多模态融合的技术,目前神经网络方法已在很多领域得到了应用,例如视觉和媒体问答[67]、手势识别[68]和视频描述生成[69]等,这些应用充分利用了神经网络方法较强的学习能力和分类性能。近期神经网络方法通过使用循环神经网络(Recurrent Neural Network,RNN)和长短期记忆网络(Long Short-Term Memory,LSTM)来融合时间多模态信息,例如文献[50]使用 LSTM 模型进行连续多模态情感识别,相对于 MKL 和 GM 方法表现出更优的性能。此外,神经网络多模态融合方法在图像字幕处理任务中表现良好,主要模型包括神经图像字幕模型[70]、多视图模型[71]等。神经网络方法在多模态融合中的优势是具备大数据学习能力,其分层方式有利于不同模态的嵌入,具有较好的可扩展性,但缺点是随着模态的增多,模型可解释性变差。

3 多模态对齐方法

多模态对齐是多模态融合的关键技术之一,指从两个或多个模态中查找实例子组件之间的对应关系。例如,给定一个图像和一个标题,需找到图像区域与标题单词或短语的对应关系[72]。多模态对齐方法分为显式对齐和隐式对齐。显式对齐关注模态之间子组件的对齐问题,而隐式对齐则是在深度学习模型训练期间对数据进行潜在对齐,如表 3 所示。

表 3 多模态对齐方法

Table 3 Multimodal alignment methods

对齐方法	对齐类型	模态类型	参考文献
显示对齐	无监督方法	视频 + 文本	文献[73-74]
		视频 + 语音	文献[75]
	监督方法	视频 + 文本	文献[76-77]
		图像 + 文本	文献[78]
隐式对齐	图像模型方法	语音/文本 + 文本	文献[79]
	神经网络方法	图像 + 文本	文献[80]
		视频 + 文本	文献[81]

3.1 显式对齐方法

无监督方法在不同模态的实例之间没有用于直接对齐的监督标签,例如:文献[73]提出的动态时间扭曲(Dynamic Time Warping,DTW)方法是一种动态规划的无监督学习对齐方法,已被广泛用于对齐多视图时间序列;文献[74]根据相同物体的外貌特征来定义视觉场景和句子之间的相似性,从而对齐电视节目和情节概要。上述两个研究成果都在没有监督信息的前提下,通过度量两个序列之间的相似性,在找到它们之间的最佳匹配后按时间对齐(或插入帧),实现字符标识和关键字与情节提要和字幕之间的对齐。还有类似 DTW 的方法用于文本、语音和视频的多模态对齐任务,例如文献[75]使用动态贝叶斯网络将扬声器输出语音与视频进行对齐。尽管无监督对齐方法无需标注数据,可以节省数据标注成本,但对实例的规范性要求较高,需具备时间一致性且时间上没有较大的跳跃和单调性,否则对齐性能会急剧下降。

监督方法是从无监督的序列对齐技术中得到启发,并通过增强模型的监督信息来获得更好的性能,通常可以将上述无监督方法进行适当优化后直接用于模态对齐。该方法旨在不降低性能的前提下,尽量减少监督信息,即弱监督对齐。例如:文献[76]提出一种类似于规范时间扭曲的方法,主要利用现有(弱)

监督对齐数据完成模型训练,从而提升深度学习模型性能;文献[77]利用少量监督信息在图像区域和短语之间寻找协调空间进行对齐;文献[78]训练高斯混合模型,并与无监督的潜变量图像模型同时进行弱监督聚类学习,使音频信道中的语音与视频中的位置及时对齐。因此,监督方法的对齐性能总体上优于无监督方法,但需要以标注数据为基础,而准确把握监督信息的参与程度是一项极具挑战的工作。

3.2　隐式对齐方法

图像模型方法最早用于对齐多种语言之间的语言机器翻译及语音音素的转录[79],即将音素映射到声学特征生成语音模型,并在模型训练期间对语音和音素数据进行潜在对齐。构建图像模型需要大量训练数据或手工运行,因此随着深度学习研究的深入及训练数据的有限,该方法已不适用。

神经网络方法是目前解决机器翻译问题的主流方法,无论是使用编解码器模型还是通过跨模态检索都表现出较好的性能。利用神经网络模型进行模态隐式对齐,主要是在模型训练期间引入对齐机制,通常会考虑注意力机制。例如,图像自动标注应用中在生成连续单词时[80],注意力机制允许解码器(通常是 RNN)集中在图像的特定部分,该注意力模块为一个浅层神经网络,其与目标任务一起完成端到端训练。该方法目前已被广泛应用于语音数据标注、视频文本对齐和视频转录等领域[81],但由于深度神经网络的复杂性,因此设计注意力模块具有一定的难度。

4　公开数据集

多模态融合技术作为一个具有极大发展潜力的研究方向,大量研究人员一直对现有模型进行不断创新和探索以完善数据集,提升多模态深度学习模型性能,提高预测准确率。表 4 列举了常见用于多模态融合技术研究和应用的公开数据集,并给出各数据集目前的最优学习结果,其中包括准确率(Accurary,ACC)、正确分类率(Correct Classification Rate,CCR)、等错误率(Equal Error Rate,EER)和平均精度均值(Mean Average Precision,MAP)。

表 4　多模态融合公开数据集

Table 4　Open datasets for multimodal fusion

数据集名称	模态类型	应用领域	参考文献	最优结果
UTD-MHAD	深度和惯性的传感器数据	人类行为识别	文献[82-83]	ACC 为 95.38%
ChaLearn	语音和骨骼姿势数据	人类行为识别	文献[84]	ACC 为 85%
Berkeley MHAD	多视点深度图像和骨骼姿势数据	人类行为识别	文献[85-86]	CCR 为 97.6% ACC 为 99.8%
MegaFace	人脸图片 4.7M 张,每张图片包含 100 个脸	人脸识别	文献[87]	ACC 为 86.47%
H-MOG	9 部智能手机传感器和交互数据	手机持续认证	文献[88]	ACC 为 92.1%
RECOLA	音频、视觉和生理数据	情感识别	文献[89-90]	ACC 为 65%
M2VTS	声音信号和视频信号	图像问答	文献[91]	ACC 为 96.57%
Pinterest Multimodal	图像和文本数据	多模文字嵌入	文献[92]	ACC 为 97.6% EER 为 0.97
TULIPS1	声音信号和视频信号	视听语音识别	文献[93]	EER 为 1.74
FCVID	视频和音频数据	视频分类	文献[12]	ACC 为 95.21% EER 为 1.5
Wikipedia	标有语义类别的文本图像对文档集	机器翻译	文献[94]	MAP 为 0.360 8
KinectFaceDB	深度图像和面部标注数据	面部识别	文献[95-96]	ACC 为 87.63%
NUS-WIDE	图片和标签(含独特标签)	跨媒体检索	文献[97]	MAP 为 0.365

5　多模态融合技术研究展望

现有多模态融合技术可有效提升深度学习模型性能,但仍有一些问题亟待解决,例如跨模态迁移学习、特征间语义鸿沟、模态泛化能力等。

1)多模态融合技术在深度学习等新兴研究领域的进一步应用探索。随着深度学习应用的不断深入,多模态融合技术的优势凸显,如基于传感器数据、人类活动识别、医学研究等多模态融合方面,这些领域会在未来几年获得更多的关注。特别是自主机器人和多媒体两个应用领域中的多模态融合问题正在引起深度学习研究人员的极大关注,例如视频转录、图像字幕、在线聊天机器人等。

2)多模态融合技术为多数据集之间的跨模态迁移学习提供了桥梁,尽管迁移学习已广泛应用于多模态深度学习领域,但由于长期以来人工数据标注成本高和许多领域的标注数据资源稀缺问题,因此基于多模态融合的迁移学习仍是下一步将重点关注的方向。

3)目前深度学习多模态融合中的语义冲突、重复和噪声等问题仍未得到较好解决。虽然注意力机制可以部分处理这些问题,但其主要为隐式运行,不易受到

主动控制。解决该问题的一种有效方法为将逻辑推理能力集成到多模态融合技术中,深度学习与逻辑推理的结合将赋予机器智能更多的认知能力。

4)多模态融合技术将在情感识别与分析领域发挥更大作用。目前利用多模态融合进行情感识别研究仍处于部分融合阶段,尚未建立一个情感分析的综合数据库,下一步可将人体的所有特征包括面部表情、瞳孔扩张、语言、身体运动、体温等进行多模态融合,以获得更全面、详细的情感识别结果。

5)多模态融合中的特征间语义鸿沟、模态泛化能力、多模态组合评价标准等关键问题仍将得到持续关注。为解决多模态特征的语义鸿沟,实现各模态信息的无障碍交流互通,需要探索更有效的语义嵌入方法。模态泛化能力是将已有模态上学习的多模态表示和模型推广到未知模态上,使机器具备高效、准确学习数据库外数据的能力。如何高效、规范地组合模态是一个从理论到具体算法都亟待解决的问题,并且还需设计一个更具普适性的评价标准来判定组合形式的优劣。

6)多模态深度学习的目标函数通常为非凸优化函数,目前的深度学习训练算法不能有效避开鞍点,导致寻优过程失败,使得研究人员无法获知是优化过程未找到最优解导致预测结果较差,还是其他模态融合和模态对齐中存在问题。针对该情况,需设计求解非凸优化问题的求解算法。

6 结束语

本文总结了深度学习领域多模态融合技术的研究现状,对融合架构、融合方法、模态对齐等进行重点分析。融合架构按照特征融合方式的不同,分为联合架构、协同架构和编解码器架构。融合方法包括早期、晚期、混合这 3 种与模型无关的方法以及多核学习、图像模型这 2 种基于模型的方法。模态对齐是多模态融合技术的难点,其常用处理方式为显示对齐和隐式对齐。近期在模态融合技术上的研究促进了大量新型多模态算法的提出,并且拓展了多模态学习的应用范围。这些模型和算法各有优缺点,可在不同领域应用中发挥优势和作用。多模态深度学习作为一种能使机器具有更多人类智能特性的技术,有望在今后获得长足发展。后续将针对模态语义冲突消解、多模态组合评价、跨模态转移学习等问题进行深入研究,促进多模态融合技术在深度学习等新兴领域的应用与发展。

参考文献

[1] RAMACHANDRAM D,TAYLOR G W. Deep multimodal learning:a survey on recent advances and trends[J]. IEEE Signal Processing Magazine,2017,34(6):96-108.

[2] BALTRUSAITIS T,AHUJA C,MORENCY L P. Multimodal machine learning: a survey and taxonomy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019,41(2):423-443.

[3] PENG Yuxin,QI Jinwei. CM-GANs:cross-modal generative adversarial networks for common representation learning[J]. Multimedia,2019,15(1):1-13.

[4] LEDERER C,ALTSTADT S,ANDRIAMONJE S,et al. A new approach to cross-modal multimedia retrieval[C]// Proceedings of the 18th ACM International Conference on Multimedia. New York,USA:ACM Press,2010:251-260.

[5] LIU Yanan, FENG Xiaoqing, ZHOU Zhiguang. Multimodal video classification with stacked contractive autoencoders [J]. Signal Processing, 2016, 120 (1): 761-766.

[6] WU S, BONDUGULA S, LUISIER F. Zeroshot event detection using multi-modal fusion of weakly supervised concepts [C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA:IEEE Press,2014:2665-2672.

[7] HABIBIAN A, MENSINK T, SNOEK C G M. Video2vec embeddings recognize events when examples are scarce[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(10):2089-2103.

[8] LI Xia,LU Guanming,YAN Jingjie,et al. A review of multimodal dimension emotion prediction[J]. Journal of Automation,2018,44(12):2142-2159. (in Chinese)
李霞,卢官明,闫静杰,等. 多模态维度情感预测综述[J]. 自动化学报,2018,44(12):2142-2159.

[9] XIE Zhibing, GUAN Ling. Multimodal information fusion of audio emotion recognition based on kernel entropy component analysis[J]. International Journal of Semantic Computing,2013,7(1):25-42.

[10] QI Jinwei, PENG Yuxin, YUAN Yuxin. Cross-modal bidirectional translation via reinforcement learning[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: [s. n.], 2018:2630-2636.

[11] LECUN Y,BENGIO Y,HINTON G. Deep learning[J]. Nature,2015,521(7553):436-451.

[12] JIANG Yugang,WU Zuxuan,WANG Jun,et al. Exploiting feature and class relationships in video categorization with regularized deep neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,40(2): 352-364.

[13] PORIA S, CAMBRIA E, HOWARD N, et al. Fusing audio, visual and textual clues for sentiment analysis from multimodal content [J]. Neurocomputing, 2016, 174:50-59.

[14] ZADEH A, LIANG P P, PORIA S, et al. Multi-attention recurrent network for human communication comprehension [C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2018:1-35.

[15] FUKUI A, PARK D H, YANG D, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding [C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Palo Alto,USA:AAAI Press,2016:457-468.

[16] LU J S, YANG J W, BATRA D, et al. Hierarchical question-image co-attention for visual question answering[C]//Proceedings of the 30th Conference on Neural Information Processing Systems. Barcelona, Spain:[s. n.],2016:289-297.

[17] PANG L, NGO C W. Mutlimodal learning with deep Boltzmann machine for emotion prediction in user generated videos [C]//Proceedings of the 5th Asian Conference on Machine Learning. New York, USA: ACM Press,2015:619-622.

[18] HUANG J, KINGSBURY B. Audio-visual deep learning for noise robust speech recognition[C]//Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press,2013:7596-7599.

[19] WANG Bokun, YANG Yang, XU Xing, et al. Adversarial cross-modal retrieval[C]//Proceedings of 2017 ACM Multimedia Conference. New York, USA: ACM Press,2017:154-162.

[20] PENG Yuxin, QI Jinwei, YUAN Yuxi. Modality-specific cross-modal similarity measurement with recurrent attention network [J]. IEEE Transactions on Image Processing,2018,27(11):5585-5599.

[21] SOCHER R, KARPATHY Q V L A, MANNING C D, et al. Grounded compositional semantics for finding and describing images with sentences[J]. Transactions of the Association for Computational Linguistics,2014,2(1): 207-218.

[22] PAN Yingwei, MEI Tao, YAO Ting, et al. Jointly modeling embedding and translation to bridge video and language[C]//Proceedings of 2016 Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press,2016:4594-4602.

[23] KIROS R, SALAKHUTDINOV R, RICHARD S Z. Unifying visual-semantic embeddings with multimodal neural language models[J]. Computer Science, 2014, 14(11):2953-2968.

[24] HUANG Xin, PENG Yuxin, YUAN Mingkuan. Cross-modal common representation learning by hybrid transfer network[C]//Proceedings of the 26th International Joint Conference on Artificial. Washington D. C., USA:IEEE Press,2017:1893-1900.

[25] LIANG Xiaodan, HU Zhiting, ZHANG Hao, et al. Recurrent topic-transition GAN for visual paragraph generation[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Washington D. C., USA:IEEE Press,2017:3362-3371.

[26] GAO Lianli, GUO Zhaoguo. Video captioning with attention based LSTM and semantic consistency[J]. IEEE Transactions on Multimedia,2017,19(9):2045-2055.

[27] YANG Yang, ZHOU Jie, AI Jiangbo. Video captioning by adversarial LSTM[J]. IEEE Transactions on Image Processing,2018,27(11):5600-5611.

[28] ZHANG Han, XU Tao, LI Hongsheng. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Washington D. C., USA:IEEE Press,2017:5907-5915.

[29] ZADEH A, CHEN M, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis [C]// Proceedings of EMNLP'17. Washington D. C., USA: IEEE Press,2017:1103-1114.

[30] NGIAM J, KHOSLA A, KIM M, et al. Multimodal deep learning [C]//Proceedings of the 28th International Conference on Machine Learning. Washington D. C., USA:IEEE Press,2011:689-696.

[31] SRIVASTAVA N, SALAKHUTV R. Learning representations for multimodal data with deep belief nets[C]//Proceedings of International Conference on Machine Learning. Washington D. C., USA:IEEE Press,2012:1-8.

[32] HE Yonghao, XIANG Shiming, KANG Cuicui, et al. Cross-modal retrieval via deep and bidirectional representation learning [J]. IEEE Transactions on Multimedia,2016,18(7):1363-1377.

[33] LEDERER C, ALTSTADT S, ANDRIAMONJE S, et al. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework [C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Palo Alto, AAAI Press, 2015: 2346-2352.

[34] LIONG V E, LU J W, TAN Y P, et al. Deep coupled metric learning for cross-modal matching [J]. IEEE Transactions on Multimedia,2017,19(6):1234-1244.

[35] PENG Yuxin, QI Jinwei, HUANG Xin. CCL: cross-modal correlation learning with multigrained fusion by hierarchical network [J]. IEEE Transactions on Multimedia, 2017, 20(2):405-420.

[36] MOR N, WOLF L, POLYAK A, et al. A universal music translation network[J]. Statistics,2018,2(3):1-14.

[37] HUANG X, LIU M Y, BELONGIE S, et al. Multimodal unsupervised image-to-image translation[C]//Proceedings of the 15th European Conference on Computer Vision. Berlin, Germany:Springer,2018:172-189.

[38] DMELLO S K, KORY J. A review and meta-analysis of multimodal affect detection systems[J]. ACM Computing Surveys,2015,47(3):43-50.

[39] ZENG Z, PANTIC M, ROISMAN G I, et al. A survey of affect recognition methods: audio, visual, and spontaneous expressions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2009,31(1):39-58.

[40] CASTELLANO G, KESSOUS L, CARIDAKIS G. Emotion recognition through multiple modalities: face, body gesture, speech[J]. Affect and Emotion in Human-Computer Interaction,2008,4868(1):92-103.

[41] RAMIREZ G A, BALTRUSAITIS T, MORENCY L P. Modeling latent discriminative dynamic of multi-dimensional affective signals [C]//Proceedings of International Conference on Affective Computing and Intelligent Interaction. Berlin, Germany: Springer, 2011: 396-406.

[42] LAN Z Z, LEI B, YU S I, et al. Multimedia classification and event detection using double fusion [J]. Multimedia Tools and Applications,2014,71(1):333-347.

[43] BUCAK S S, JIN R, JAIN A K. Multiple kernel learning for visual object recognition: a review [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2014,36(7):1354-1369.

[44] JAQUES N, TAYLOR S, SANO A, et al. Multi-kernel learning for estimating individual wellbeing [C]// Proceedings of NIPS Workshop on Multimodal Machine

Learning. Montreal, Quebec: [s. n.], 2015: 1-7.
[45] SIKKA K, DYKSTRA K, SATHYANARAYANA S, et al. Multiple kernel learning for emotion recognition in the wild[C]//Proceedings of the 15th ACM International Conference on Multimodal Interaction. New York, USA: ACM Press, 2013: 517-524.
[46] GURBAN M, THIRAN J P, DRUGMAN T, et al. Dynamic modality weighting for multi-stream HMMs in audio-visual speech recognition[C]//Proceedings of the 16th International Conference on Multimodal Interfaces. Istanbul, Turkey: [s. n.], 2013: 237-240.
[47] BALTRUSAITIS T, BANDA N, ROBINSON P. Dimensional affect recognition using continuous conditional random fields [C]//Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face, Gesture Recognition. Washington D. C., USA: IEEE Press, 2013: 1-8.
[48] JIANG Xinyan, WU Fei, ZHANG Yin, et al. The classification of multi-modal data with hidden conditional random field [J]. Pattern Recognition Letters, 2015, 51(6): 63-69.
[49] KAHOU S E, BOUTHILLIER X, LAMBLIN P, et al. EmoNets: multimodal deep learning approaches for emotion recognition in video[J]. Journal on Multimodal User Interfaces, 2016, 10(2): 99-111.
[50] WOLLMER M, METALLINOU A, EYBEN F, et al. Context sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling [C]//Proceedings of the 11th Annual Conference of the International Speech Communication Association. Makuhari, Japan: [s. n.], 2010: 2362-2365.
[51] CHEN Shizhe, JIN Qin. Multi-modal dimensional emotion recognition using recurrent neural networks [C]// Proceedings of the 5th International Workshop on Audio/ Visual Emotion Challenge. New York, USA: ACM Press, 2015: 49-56.
[52] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.
[53] MARTINEZ H P, YANNAKAKIS G N. Deep multimodal fusion [C]//Proceedings of the 16th International Conference on Multimodal Interaction. Istanbul, Turkey: [s. n.], 2014: 34-41.
[54] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Proceedings of Advances in Neural Information Processing Systems. Berlin, Germany: Springer, 2014: 568-576.
[55] ROBIN R, MURPH Y. Computer vision and machine learning in science fiction[J]. Science Robotics, 2019, 4(30): 7221-7235.
[56] KAHOU S E, PAL C, BOUTHILLIER X, et al. Combining modality specific deep neural networks for emotion recognition in video[C]//Proceedings of the 15th ACM on International Conference on Multimodal Interaction. New York, USA: ACM Press, 2013: 543-550.
[57] WU D, PIGOU L, KINDERMANS P J, et al. Deep dynamic neural networks for multimodal gesture segmentation and recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(8): 1583-1597.
[58] GONEN M, ALPAYDN E. Multiple kernel learning algorithms[J]. Journal of Machine Learning Research, 2011, 12(3): 2211-2268.
[59] YEH Y R, LIN T C, CHUNG Y Y, et al. A novel multiple kernel learning framework for heterogeneous feature fusion and variable selection[J]. IEEE Transactions on Multimedia, 2012, 14(3): 563-574.
[60] MCFEE B, LANCKRIET G R G. Learning multi-modal similarity[J]. Journal of Machine Learning Research, 2011, 12(3): 491-523.
[61] LIU Fayao, ZHOU Luping, SHEN Chunhua, et al. Multiple kernel learning in the primal for multimodal Alzheimer's disease classification[J]. IEEE Journal of Biomedical and Health Informatics, 2014, 18(3): 984-990.
[62] SUTTON C, MCCALLUM A. Introduction to conditional random fields for relational learning[M]//GETOOR L, TASKAR B. Introduction to statistical relational learning. Cambridge, USA: MIT Press, 2006: 93-127.
[63] FIDLER S, SHARMA A, URTASUN R. A sentence is worth a thousand pixels holistic CRF model[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2013: 1995-2002.
[64] REITER S, SCHULLER B, RIGOLL G. Hidden conditional random fields for meeting segmentation[C]//Proceedings of IEEE International Conference on Multimedia and Expo. Washington D. C., USA: IEEE Press, 2007: 639-642.
[65] SONG Y, MORENCY L P, DAVIS R. Multimodal human behavior analysis: learning correlation and interaction across modalities [C]//Proceedings of the 14th International Conference on Multimodal Interaction. Washington D. C., USA: IEEE Press, 2012: 27-30.
[66] SONG Y L, MORENCY L P, DAVIS R. Multi-view latent variable discriminative models for action recognition[C]//Proceedings of the 14th International Conference Multimodal Interaction. Washington D. C., USA: IEEE Press, 2012: 2120-2127.
[67] GAO Haoyuan, MAO Junhua, ZHOU Jie, et al. Are you talking to a machine dataset and methods for multilingual image question answering[C]//Proceedings of the 29th Annual Conference on Neural Information Processing Systems. Berlin, Germany: Springer, 2015: 2296-2304.
[68] NEVEROVA N, WOLF C, TAYLOR G, et al. ModDrop: adaptive multi-modal gesture recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(8): 1692-1706.
[69] JIN Qin, LIANG Junwei. Video description generation using audio and visual cues[C]//Proceedings of the 5th ACM International Conference on Multimedia Retrieval. New York, USA: ACM Press, 2016: 239-242.
[70] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: a neural image caption generator[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 3156-3164.
[71] RAJAGOPALAN S S, MORENCY L P, BALTRUSAITIS T, et al. Extending long short-term memory for multi-view structured learning[C]//Proceedings of the 14th European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 338-353.

[72] ANDREJ K,LI F F. Deep visual-semantic alignments for generating image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(4):664-676.

[73] TAPASWI M,AUML M B,STIEFELHA R. Aligning plot synopses to videos for story-based retrieval [J]. International Journal of Multimedia Information Retrieval,2015,4(1):3-16.

[74] TAPASWI M, Auml M B, STIEFELHA R. Book2Movie: aligning video scenes with book chapters[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 1827-1835.

[75] TRIGEORGIS G,NICOLAOU M A,ZAFEIRIOU S, et al. Deep canonical time warping[C]//Proceedings of Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press,2016:5110-5118.

[76] PIOTR B, RÉMI L, EDOUARD G, et al. Weakly-supervised alignment of video with text [C]// Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA:IEEE Press, 2015:4462-4470.

[77] ZHUY K,KIROS R,ZEMEL R, et al. Aligning books and movies: towards story-like visual explanations by watching movies and reading books[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA:IEEE Press,2015:19-27.

[78] MAO Junhua,HUANG J,TOSHEV A,et al. Generation and comprehension of unambiguous object descriptions [C]//Proceedings of Computer Vision and Pattern Recognition Conference. Washington D. C., USA:IEEE Press,2016:11-20.

[79] DENGY G,BYRNE W. HMM word and phrase alignment for statistical machine translation [C]//Proceedings of Conference on Human Language Technology and Empirical Methods in Natural. New York, USA: ACM Press, 2005: 169-176.

[80] KELVIN X,JIMMY B,RYAN K,et al. Show, attend and tell: neural image caption generation with visual attention[C]//Proceedings of the 32nd International Conference on Machine Learning. New York, USA: ACM Press,2015:2048-2057.

[81] YU Haonan, WANG Jiang, HUANG Zhiheng. Video paragraph captioning using hierarchical recurrent neural networks[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA:IEEE Press,2016:4584-4593.

[82] CHEN C,JAFARI R,KEHTARNAV N. UTD-MHAD: a multimodal data set for human action recognition utilizing a depth camera and a wearable inertial sensor[C]// Proceedings of 2015 IEEE International Conference on Image Processing. Washington D. C., USA:IEEE Press, 2015:168-172.

[83] KHAIRE P, IMRAN J, KUMAR P. Human activity recognition by fusion of RGB, depth, and skeletal data[C]//Proceedings of the 2nd International Conference on Computer Vision and Image Processing. Washington D. C., USA:IEEE Press,2018:409-421.

[84] ESCALERA S,BARÓ X,GONZÀLEZ J. ChaLearn looking at people challenge 2014: dataset and results [C]// Proceedings of European Conference on Computer Vision. Berlin,Germany:Springer,2015:459-473.

[85] OFLI F,CHAUDHRY R,KURILLO G,et al. Berkeley MHAD: a comprehensive multimodal human action databases [C]//Proceedings of 2013 IEEE Workshop on Applications of Computer Vision. Washington D. C., USA:IEEE Press,2013:53-60.

[86] NG H,VUN T T Y,TONG H L,et al. Action classification on the Berkeley multimodal human action dataset[J]. Journal of Engineering and Applied Sciences, 2017, 12(3):520-526.

[87] WANG Mei,DENG Weihong. Deep face recognition: a survey [EB/OL]. [2020-01-07]. https://arxiv.org/abs/1804.06655.

[88] SITOVA Z,SEDENKA J,YANG Q,et al. HMOG: new behavioral biometric features for continuous authentication of smartphone users[J]. IEEE Transactions on Information Forensics Security,2016,11(5):877-892.

[89] RINGEVAL F, SONDEREGGER A, SAUER J, et al. Introducing the Recola multimodal corpus of remote collaborative and affective interactions[C]//Proceedings of the 10th IEEE International Conference on Automatic Face and Gesture Recognition. Washington D. C., USA: IEEE Press,2013:1-8.

[90] CHENG Yanfen, CHEN Yaoxin, CHEN Yiling, et al. Speech emotion recognition with attention mechanism and hierarchical context [J]. Journal of Harbin University of Technology, 2019, 51(11): 100-107. (in Chinese)
程艳芬,陈垚鑫,陈逸灵,等.嵌入注意力机制并结合层级上下文的语音情感识别[J].哈尔滨工业大学学报,2019,51(11):100-107.

[91] WU Q, TENEY D, WANG P, et al. Visual question answering: a survey of methods and datasets [J]. Computer Vision and Image Understanding, 2017, 163: 21-40.

[92] MAO Junhua, XU Jiajing, JING Yushi, et al. Training and evaluating multimodal word embeddings with large-scale Web annotated images [C]//Proceedings of Advances in Neural Information Processing Systems. Barcelona, Spain: [s. n.], 2016: 442-450.

[93] SEONG T W,IBRAHIM M Z. A review of audio-visual speech recognition[J]. Journal of Telecommunication, Electronic and Computer Engineering, 2018, 10(1): 35-40.

[94] GEHRING J,AULI M,GRANGIER D,et al. Convolutional sequence to sequence learning[C]//Proceedings of the 34th International Conference on Machine Learning. New York, USA:ACM Press,2017:1243-1252.

[95] MIN R,KOSE N,DUGELAY J L. KinectFaceDB: a Kinect database for face recognition[J]. IEEE Transactions on Systems Man and Cybernetics,2014,44(11):1534-1548.

[96] WEI Wei, JIA Qingxuan. 3D Facial expression recognition based on Kinect[J]. International Journal of Innovative Computing Information and Control, 2017, 13(6):1843-1854.

[97] PENG Yuxin, HUANG Xin, QI Jinwei. Cross-media shared representation by hierarchical learning with multiple deep networks [C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Press, 2016: 3846-3853.

编辑　陆燕菲