计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331,CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目：多模态深度学习综述
作者：孙影影，贾振堂，朱昊宇
网络首发日期：2020-09-30
引用格式：孙影影，贾振堂，朱昊宇．多模态深度学习综述．计算机工程与应用.
https://kns.cnki.net/kcms/detail/11.2127.TP.20200929.1838.002.html

网络首发时间：2020-09-30 10:08:58
网络首发地址：https://kns.cnki.net/kcms/detail/11.2127.TP.20200929.1838.002.html

多模态深度学习综述

孙影影，贾振堂，朱昊宇
上海电力大学 电子与信息工程学院，上海 200090

摘 要：模态是指人接收信息的方式，包括听觉、视觉、嗅觉、触觉等多种方式。多模态学习是指通过利用多模态之间的互补性，剔除模态间的冗余性，从而学习到更好的特征表示。多模态学习的目的是建立能够处理和关联来自多种模式的信息的模型，它是一个充满活力的多学科领域，具有日益重要和巨大的潜力。目前比较热门的研究方向是图像、视频、音频、文本之间的多模态学习。着重介绍了多模态在视听语音识别、图文情感分析、协同标注等实际层面的应用，以及在匹配和分类、对齐表示学习等核心层面的应用，并针对多模态学习的核心问题：匹配和分类、对齐表示学习方面给出了说明。最后对多模态学习中常用的数据集进行了介绍，并展望了未来多模态学习的发展趋势。

关键词：多模态学习；多模态应用；多模态融合；共享表示空间
文献标志码：A 中图分类号：TP181 doi：10.3778/j.issn.1002-8331.2002-0342

孙影影，贾振堂，朱昊宇. 多模态深度学习综述.计算机工程与应用
SUN Yingying, JIA Zhentang, ZHU Haoyu. Survey of Multimodal Deep Learning.Computer Engineering and Applications

Survey of Multimodal Deep Learning

SUN Yingying, JIA Zhentang, ZHU Haoyu
College of Electronics and Information Engineering, Shanghai University of Electric Power, Shanghai 200090, China

Abstract: Modal refers to the way people receive information, including hearing, vision, smell, touch and other ways. Multimodal learning refers to learning better feature representation by using the complementarity between Multimodes and eliminating the redundancy between them. The purpose of multimodal learning is to build a model that can deal with and correlate information from multiple modes. It is a dynamic multidisciplinary field, with increasing importance and great potential. At present, the popular research direction is multimodal learning among image, video, audio and text. This paper focuses on the application of multimodality in audio-visual speech recognition, image and text emotion analysis, collaborative annotation and other practical levels, as well as the application in the core level of matching and classification, alignment representation learning, and gives an explanation for the core issues of multimodal learning: matching and classification, alignment representation learning. Finally, the common data sets in multimodal learning are introduced, and the development trend of multimodal learning in the future is prospected.

Keyword: Multimodal learning；Multimodal application；Multimodal fusion；Shared representation space

1 引言

每一种信息的来源都可以称为一种模态，即模态是指人接受信息的方式，人有听觉、视觉、嗅觉、触觉等多种感知方式来认识事物，当某一种模态信息缺失时，能否准确的认知事物是我们关注的重点。由于多媒体数据通常是多种信息的传递媒介，例如一段视频中会同时含有文字信息、视觉信息和听觉信息，多模态学习已成为多媒体内容分析与理解的主要手段。随着深度学习的发展，诸如图像、文本、声音、视频等多媒体数据的急剧增长，催生出对图像文本对、图像声音对等多模态问题的研究。多模态学习由来自不同模态的信息组成，一般都是包含两个或两个以上的模态，旨在联合表示不同模态的数据,捕捉不同模态之间的内在关联，实现各个模态的信息相互转化，即使在某些模态缺失的情况下能够填充在传递过程中缺少的信息。多模态深度学习给机器学习带来了巨大的机遇与挑战，文献[1]将多模态深度学习实现过程中的共有问题分为模态表示、模态传译、模态融合和模态对齐四类，并对各问题进行子分类和论述，同时列举了为解决各问题产生的神经网络模型。

通过多模态学习能够处理和关联来自多种模式

基金项目：国家自然科学基金青年科学基金批准号(No.61401269)。
作者简介：孙影影（1996-），女，硕士，主研领域：深度多模态学习，唇语识别；贾振堂（1969-），通信作者，男，硕导，副教授，主研领域：深度学习，视频监控，E-mail：462458081@qq.com；朱昊宇（1994-），男，硕士，主研领域：深度学习，故障定位。

信息的模型，对于许多实际问题，深度多模态学习常常为涉及多模式数据的问题提供了很多改进的性能。 例如，手势识别旨在理解人体的动态手势,是人机交互领域极其重要的交互方式之一，由于视频样本中手势的短期、中期以及长期时空特征性，文献[2]提出了一种基于浅三维稠密网的多模态手势识别方法，所提出的方法在手势识别公开数据集大规模离散手势数据集上进行了评估,并取得了目前最好效果。多模态学习区别于传统机器学习方法的特点可体现在下表 1 中分析[3]。

表 1　多模态学习与传统机器学习方法比较

比较层面	多模态学习方法	传统机器学习方法
融合方式	多样化	较单一
对 *GPU* 的要求	较高	不高
特征学习方式	数据中学习特征	手动设计学习特征
网络中超参数数目	丰富	较少
输入数据预处理要求	不严格	严格
融合方法和模态数量扩展程度	易扩展	不易扩展
实验效果	较准确	不准确

多模态研究支持计算机视觉领域的许多关键应用，如图像字幕，跨模态检索。由于许多多模态数据具有显著的弱配对特性，即模态之间没有样本到样本的对应关系，而是一种模态中的样本类别对应另一种模态中的样本类别。这为跨模式的检索学习提供了巨大的挑战[4]。本文的目的是针对深度多模态学习的几个应用方面进行的介绍，进而指出多模态学习的本质问题。由于近些年来在主要会议和期刊上发表此类文章的数量的不断增加，我们更加坚信了多模态深度学习的广大应用前景。如下图 1 为深度多模态学习论文的发表数量，涉及到计算机科学、工程方面的文章。

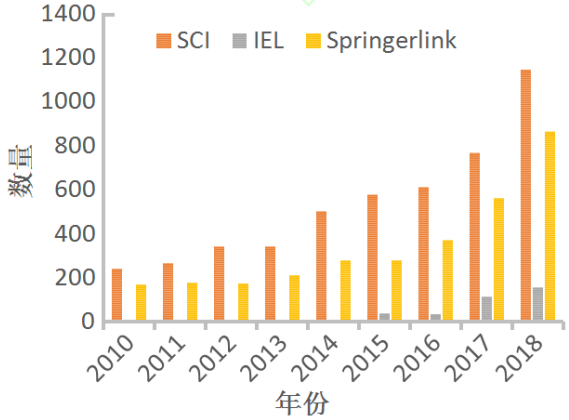

图 1　多模态学习论文的发表情况

从图表增长趋势可以看出近年来有关多模态学习的论文数量增长较快，并有持续增长的趋势。重要的原因是多模态涉及到的领域的普遍化，例如，在图像配准问题的处理上，由于图像灰度特性的非线性变化，如何对多模态图像有效测量，就需要依靠多模态深度学习来分析图像特征[5]；医疗上综合反映研究对象的生物学信息仍然是一种迫切的需求和重大的挑战，多模态核磁共振图像联合分析为揭开脑结构变化和脑功能变化的关系提供了可能,另外这种联合分析提供的全面医学影像信息对探索脑工作原理具有重要作用[6]；在航天领域研究太阳射电频谱时，由于频率的多样性，需要将不同频率信道捕获的太阳射电频谱看作不同的模态，学习这些不同频率信道的太阳射电频谱也是多模态学习的范畴[7]。多模态深度学习作为机器学习的最新发展，其研究成果在军事、农业、医学、安防等诸多场景都具有重要的应用前景。作为一种能让机器拥有更多人类智能特性的学习方法，多模态深度学习定能在之后的一个时期获得长足的发展。

本文从计算机视觉领域着手介绍多模态深度学习，第一节介绍多模态学习的发展过程；第二节简单介绍了多模态学习的几个主要研究方向；第三节针对实现多模态的关键细节进行说明比较；第四节说明多模态未来的发展趋势。

2　应用发展过程

1956 年，心理学家 Frank Rosenblatt 首次提出了可以模拟人类感知能力的机器，并称之为感知机（Perceptron），感知机是有单层计算单元的神经网络。由于单层感知机的局限性，后来有了多层感知机，但由于多层感知机对隐藏层权值的训练存在问题，有很长一段时期神经网络的发展进入了瓶颈[8]。最经典的发展就是反向传播神经网络（Backpropagation algorithm,BP），这是一种监督学习算法，为以后的多模态深度学习打下了坚实的基础。2006 年，Geoffrey Hinton 提出多层人工神经网络模型有很强的学习能力，深度学习模型可以学习到原始数据更本质的表示，且对于深度神经网络很难训练到最优问题，提出了逐层训练的方法[9]。

多模态学习作为深度学习的一种，最早始于 1970 年，经历了几个发展阶段，在 2010 年后全面步入深度学习阶段。最早的多模态研究应用之一是视听语音识别，这一点在 McGurk 效应中首次得到证明，大多数受试者都将带有语音 ba 和视觉 ga 视为 da，这是由于在语音感知过程中听觉和视觉之间的相互作用而产生的结果[10]，这些结果促使许多研

究人员将他们的研究领域扩展到视觉信息上。于是在进行声音识别过程中，研究人员开始联合视频和声音两个模态，结果比在原来的只有单个声音模态输入的系统上实现了较大的飞跃，多模态机器学习开始表现出其优秀的学习能力。

另一种重要的多模态应用是对多媒体数据内容的检索，从 1990 年开始，随着信息领域的发展，多媒体数据所占的比例越来越大，网络信息不再只是单纯的文字信息，图形图像、视频、声音等多媒体信息在因特网中所占比重越来越大[11]。但多媒体数据的检索技术还远远跟不上多媒体数据的迅速产生，这在一定程度上影响了多媒体信息检索技术的应用和推广。于是研究人员开始关注对多媒体内容的检索，基于内容的检索已经成为多媒体领域研究的热点。

第三类应用是在 21 世纪初围绕着新兴的多模式交互领域建立起来的，目的是了解人类的多模态行为。在计算机视觉领域，单项生物特征识别技术已经不能满足客户的要求了，多模态的解决办法被业内专家提出，并成为众多计算机视觉公司逐步去落实的事情[12]。例如虹膜识别，识别距离与人脸识别技术同时工作，两种技术实现了真正意义上的融合。由于在自动人脸检测、面部标志检测和面部表情方面取得了很大的进步，情感识别和情感计算领域在 2010 年初开始蓬勃发展。

最具代表性的应用程序之一是图像描述，它类似于给定一幅图片来获取它的文字表述。图像描述自动生成是一个融合计算机视觉、自然语言处理和机器学习的综合性问题，图像描述自动生成的主要挑战是如何评估预测描述的质量，该任务不仅需要利用模型去理解图片的内容并且还需要用自然语言去表达它们之间的关系[13]。例如，通过摄像头获取到的图像或视频，结合图像描述以及语音生成技术，可以获得对眼前事物的准确描述。

3 多模态学习的应用研究

多模态学习依据模态判别的标准不同可以有多种多样的应用。例如在太阳射电爆发分类的多模式深度学习一文中，把从不同频率信道捕获的太阳射电频谱看作不同的模态，学习这些不同频率信道的太阳射电频谱也是多模态学习的范畴[7]。多模式机器学习旨在建立能够从多种模式中处理和关联信息的模型。由于数据的异质性，不同模态之间存在鸿沟，阻碍了信息间的直接交互，多模态机器学习的研究给研究者带来了一些独特的挑战。本文主要介绍多模态学习的几个主要应用研究领域，我们仅关注三种模式：自然语言、视觉信号以及语音信号。多模态学习的应用涉及许多方面，目前比较热门的研究方向包括计算机视觉领域，医疗领域以及自动驾驶汽车等。

3.1 视听语音识别

多模态研究中应用较成熟的是视听语音识别，一种融合了语音和视觉模式的深度多模学习方法。在视听语音识别中，说话人的录音和视频都可以在培训时使用。针对视听双通道的语音识别，文献[14]建立了基于隐马尔科夫（HMM）的视听融合模型，并对模型进行训练和识别，实现了视听双通道的语音识别系统。文献[15]最先联合声音和视频对两个独立的网络分别进行音频和视觉特征的训练，利用随机梯度下降算法对网络进行优化。并引入了双线性 DNN 模型，如图 2 所示。融合发生在最后一个隐藏层，可以通过双线性 DNN 模型捕捉模态中的非线性特征之间的相关性。然后保持固定的特征空间，而在这个融合空间中训练一个深的或浅的 Softmax 网络，直到达到目标。

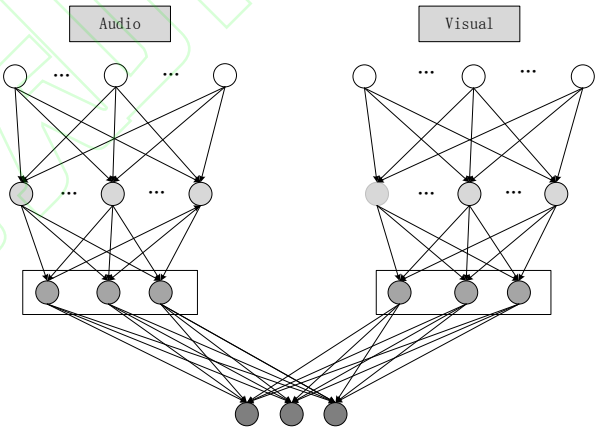

图 2　双线性 DNN

结果显示使用双线性 DNN 模型对两种模态进行训练比单一模态达到的效果好，语音识别准确度提高，但针对噪声影响较大的语音效果不好。又由于噪声因素不可避免，文献[16]以噪声环境下的自动语音识别为研究背景，建立视听信息决策层的多模态融合模型，在隐马尔科夫（HMM）统计模型的基础上，通过多模态融合处理来降低或消除音频噪声，通过 HMM 的训练步骤估计模型的参数，由关联处理最终进行融合判决，仿真结果表明应用多模态视听信息融合能有效克服噪声干扰，提高识别准确度。

3.2 图文情感分析

多模态的研究可用于学习多模态数据的情感分析，可以帮助更好地理解对某些事件的态度或观点，情感分析中的多模态数据处理一直是一项具有挑战

性的任务。首先，与传统的单一情态情感分析相比，多模态情感分析中包含着不同的表现形式，因此，情感分析方法应该有效地弥合不同模式之间的差距。

传统的情感分析方法往往不能同时考虑图片影响、特殊符号信息以及上下文信息，而导致情感分析方法准确率不高的问题，文献[17]提出了一种基于转移变量的图文融合微博情感分析方法，通过处理句子的情感从属和主题从属，引入图片因素为情感浓度来影响文本的情感分布，最后计算微博的整体情感倾向。实验结果表明，与传统情感分析模型相比，本模型测试数据集的准确率更高。由于微博文本具有长度受限、写作不规范、风格随意、主题发散等特点，针对这个问题，文献[18]提出了一种基于依存关系的情感词识别方法，通过对情感词相关依存关系的统计和分析，构建情感词识别模版以识别微博语料中的网络情感词，再利用基于点互信息量方法计算情感词的倾向性，从而构建网络情感词典。

为了挖掘不同模式下的互补信息和非冗余信息，文献[19]提出了基于视觉关注模型、语义关注模型和多模态关注模型三种模型的后期融合方案，即融合到一个多模态情感分析的整体框架中，运用了一种结合视觉注意机制的长短期记忆网络（Long Short-Term Memory，LSTM），用于捕捉图像与文本之间的关联，以达到正确获取社会图像情感的目的，模型结构如图 3。实验在 Getty image, Twitter and Flickr 三个大型数据集上对该模型的性能进行了一系列实验，结果表明，提出的方法在三个数据集上的性能优于目前最新的方法。所提出的融合模型有效地将不同的数据模式结合在一起，从而实现较理想的情感分类性能。

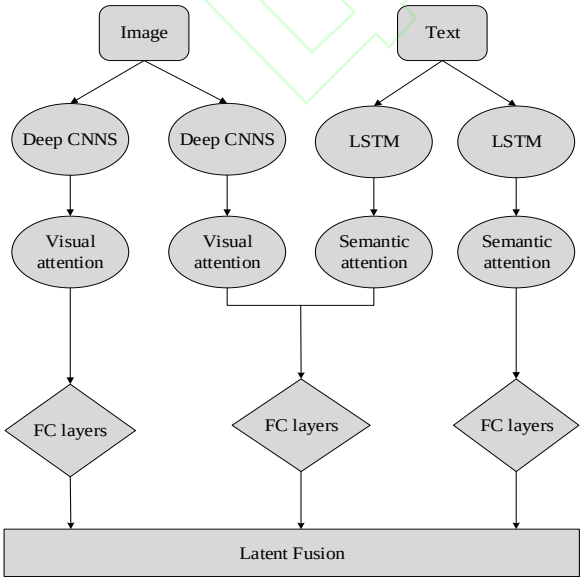

图 3　图文情感识别模型

3.3　协同标注

多模态的研究可用于多媒体数据标注，多媒体数据由文本、图像、视频、音频、时间序列等多种形式组成。有时模态数据可能会存在缺乏标注数据、样本含大量噪声以及数据收集质量不可靠等问题，可通过不同模态间的知识迁移提高质量较差模态的性能。文献[19]提出一种基于注意力机制的 LSTM 网络，利用语义一致性，捕捉视频的显著结构，探索多模态表示之间的关系来完成视频标注，但针对复杂视频信息效果不好；文献[20]在利用注意力机制的基础上，基于语言知识选择性地关注视觉属性的标注方法，该方法将神经网络中的隐藏状态映射到潜在嵌入空间，从而获得语言与视觉属性的对应关系；后来文献[21]提出一种包含属性的 LSTM 和 RNN 网络来发现图像视觉属性与语义表达之间的复杂关系，还关注了句子和视频的对应关系。文献[22]提出了一种跨模态知识迁移网络，利用源域和目标域的模式作为桥梁，将知识同时迁移到两种模态，而层共享相关子网络保留固有的跨模态语义相关性以进一步适应跨模式检索任务。事实上，不同模态的多媒体内容从各自的形式描述给定的标签，并相互补充，探索异类数据分析和多媒体注释的先进技术变得至关重要。基于这一思想，文献[23]提出了一种新的异构多媒体协同标注多模态相关学习方法，即统一空间学习，将异构媒体数据投影到一个统一的空间中，所提出的投影空间如图 4 所示。

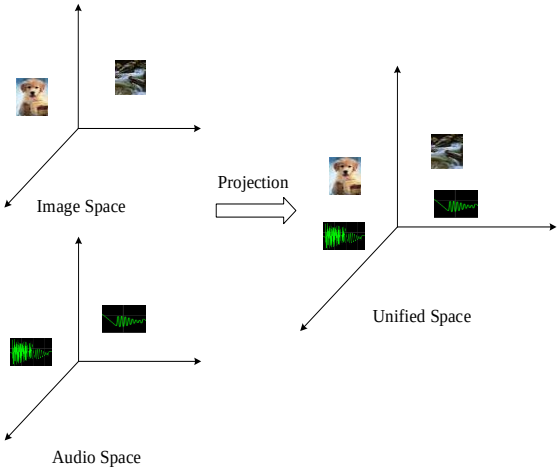

图 4　统一空间映射模型

将多媒体标注任务转化为半监督学习框架，学习不同媒体类型的不同投影矩阵。对于一个新的媒体样本，我们可以很容易地将其嵌入到统一的空间中，然后将其相邻的相关标签分配给该样本[24]。通过对图像、音频片段、视频和三维模型数据集的实验结果表明，不同的媒体内容相互协调，共同为给

定的语义标签提供了一个更为互补的轮廓，可以学习到异构媒体数据的更有效表示[25]。

3.4 匹配和分类

多模态的研究可学习图像和文本之间的共享表示特征，用于多模态的匹配和分类，匹配即特征嵌入问题，分类即预测类标签。与目前仅关注多模式匹配或分类的方法不同，文献[23]提出了一个统一的网络来共同学习图像和文本之间的多模态匹配和分类。所提出的多模态匹配和分类网络模型涉及视觉和语言之间，它可以无缝集成匹配和分类组件。其中实现两个组件的融合是关键，这就涉及到多模态融合问题。多模态信息的融合能获得更全面的特征，提高模型鲁棒性，并且保证模型在某些模态缺失时仍能有效工作[26]。

针对多模态融合问题，包括网络结构上的改进以及算法的优化两大方面：在网络结构方面，常用的是带注意力机制的递归神经网络，再利用注意力机制将文本与图像特征融合[27]。但是这种网络结构往往不能高度集中的表示数据，于是有了一种新型端到端的深度融合卷积神经网络，将二维与三维数据输入网络进行特征提取和融合，进而获得高度集中的特征表示，可应用于人脸表情识别[28]。在算法优化方面：新型高效的融合方法是哈希算法，它将弱监督方式提取出的多模态特征统一整合为二进制编码，从而使用核函数配合 SVM 进行分类[29]。文献[23]不仅提出了一个统一的网络结构，还提出了一种结合匹配和分类损失的多级训练算法，它可以使匹配和分类组件在一个统一的模型中更加兼容。通过四个众所周知的基实验表明，所提出的网络模型具有较好的鲁棒性，优于匹配或分类单独作用时的效果，对与匹配或分类相关的多模态任务有很好的推广应用前景。

3.5 对齐表示学习

多模态研究还可用于不同模态之间的对齐表示，可在不同模式之间传递所学的知识。对齐旨在挖掘不同模态之间的对应关系，从而促使学习到的多模态表示更加精确，并且也为多媒体检索提供更细致的检索线索[30]。在多模态的对齐学习中，常用最大边距学习方式结合局部对齐和全局对齐方法学习共同嵌入表示空间[29]。在跨模态检索方法中，模态与模态之间存在一定的数据相关性，基于判别性字典学习的跨模态检索方法可以增强来自不同类别的模态内数据的辨别能力，运用判别性字典来解释每种模态，通过标签对齐方法进一步增强跨模态数据的区分性和相关性[31]。

对齐的跨模态表示将对计算机视觉产生很大的影响，因为它们是机器感知理解模式之间关系的基本组成部分。在实际学习词、句子、图像以及图像区域的特征对齐表示时，提出了层次化多模态 LSTM 的密集视觉-语义嵌入方法，可以有效地学习词、句子、图像以及图像区域的对齐表示[32]。文献[33]设计了一个跨模态网络模型，它可以接受图像、声音或句子作为输入，并产生一个跨模式共享的通用表示。通过实验表明，深度跨模态表示法比以往的聚类 CCA 和线性回归都有很大的优势。因为所提出的网络能够学习高层次的特性，更容易跨模式对齐。但是当模态之间不匹配或者匹配程度低时，不容易学习它们的对齐表示，就需要设计一种深层跨模态对齐网络多次进行训练学习以尽可能消除模态间的不匹配问题[34]。

表示学习的目的是将被研究对象中所蕴含的语义信息抽象为实值向量，研究对象包括结构化数据以及图像、视频、语音、文本等非结构化数据[30]。最初基于模态相关性约束，出现了一种面向多模态表达的紧致哈希编码方法，该方法首先基于模态内和模态间的相关性约束，提出了一种新的深度学习模型生成哈希编码[35]。但是由于数据不是连续的，会造成部分模态数据的缺失问题，又发展了一种基于自适应相似结构正则化的部分多模态稀疏编码模型，能很好的解决数据稀疏造成的模态缺失问题[36]。

多模态学习的研究起源于人们日常生活中的许多实际问题，目的是帮助人们解决复杂度更高的问题。多模态学习应用可以很广泛，涉及计算机视觉领域、医疗领域、天文学探测方面以及自动驾驶汽车等[37]。从以上多模态深度学习的几个应用领域看，深度多模式学习的研究已取得较大的成果，有巨大的发展潜力。从最近几年的多模态应用方面的文章看，多模态学习有极好的发展前景，应用实际生活中具有重要的现实意义。越来越多的文章致力于从图像、声音、视频和文本等热门方向着手来寻求各个模态之间的互联想，逐渐形成一个以神经网络为基础的完善的理论体系结构，通过一系列基准实验证明了该结构的可实现性[38]。第三节将针对多模态学习的具体实现细节加以说明。

4 实现细节

4.1 多模态本质问题

从多模态在第二节的几个典型应用可以发现，来自不同模态的信息要想达到较好的实验效果，它们区分单一模态的关键在于如何构建一个共享表示空间，该共享表示空间可以融合来自两个或多个模态的特征，从而可以找出各个模态之间的对应关系[39]。我们研究多模态学习的目的就是通过建立共享

空间表示，学习不同模态之间的关系，最后实现模态之间的互联想。这样，同一现象的多个模态信息可以相互补充，当某一模态数据缺失时，多模态学习仍能达到很好的效果。

多模态学习有重要的现实意义，但是目前针对多模态学习的研究仍然十分有限。对于多模态学习，比较热门的研究方向包括多模态的表示学习、不同模态之间的相互转化、多模态融合、多模态对齐和共同学习等等。尽管多模态应用广泛，但其本质问题是在不同模态之间实现某种关联。本节我们主要介绍建立一个共享表示空间的两个关键过程：多模态融合和多模态对齐，并对多模态学习中常用的数据集进行介绍。

4.2 多模态融合

在多模态学习的早期就已经开始了对多模态融合的研究，这是多模态学习研究最多的方面之一，它的工作可以追溯到 25 年前[40]。首先，多模态融合从技术上讲，是将来自多种模态的信息集成在一起的概念，目的是通过分类方法来预测一个类。例如在医学领域，医生就诊更多根据图像在局部区域高层语义特征(如是否病变、病变类型等)的差异，粗粒度地判断图像的相似程度，针对现有的医学图像特征表达忽略了医学图像特有的高层语义特征,致使医学图像聚类效果不佳的问题,文献[41]提出了一种多模态医学图像聚类方法，就融合了医学图像纹理特征和特有形态学特征，并通过实验验证了该方法的有效性。可见多种模态的信息相互融合可以实现信息补充，提升预测结果的精度，提高预测模型的鲁棒性，使最后的结果更可靠。

一般的融合分为特征融合和决策融合，特征融合指网络一起提取的表达融合,之后接一个分类层；决策融合指模型组合,融合网络计算的分类得分。在此主要介绍特征融合，特征融合即输入两个模态的特征向量，输出融合后的向量，最常用的方法是拼接、按位乘、按位加。特征融合能有效提高某些算法的准确度，例如，针对单模态行人检测在光照条件较差、目标部分遮挡、目标多尺度时检测效果较差的问题,文献[42]提出了一种基于可见和红外双模态特征金字塔融合的行人检测算法，实验结果表明在KAIST数据集上的检测效果超过了目前该数据集上的最佳模型。从特征融合的结构上分，可分为早期融合，后期融合，后来又有了中间融合[43]。三种融合结构的特点如下表 2 所示。

表 2　融合结构特点

比较层面	早期融合	后期融合	中间融合
用途	分类	回归	分类、回归
结构	简单	较复杂	最复杂
融合方式	多种	较少	较少
特征维度	较高	一般	较低
实现难度	容易	较难	最难
处理数据异步能力	较弱	较强	最强
实验效果	不好	一般	较好

早期融合主要用于分类，在进行特征提取后立即集成，通常只是简单连接它们的表示，广泛出现在多模态学习任务中；晚期融合用于回归，一般在每个模块之后再执行集成，可以有效的处理数据的异步性，但实现程度较早期融合较难；中间融合用于分类回归，它结合了早、晚期融合的优点，同时模型复杂度和实现难度也增加了[44]。

从融合方法上看，又可分为基于核融合、基于图像模型和基于神经网络的方法，其中基于神经网络的融合方法是比较流行的方法。文献[45]把多模态表示分为联合表示和协同表示，联合表示是将多个单模信号合并到同一个表示空间，学习各个模态间的共享表示；协同表示是在信号投影之前强制执行一定相似性约束来协调它们[46]。

以图像、文本的融合为例，(X_i, Y_i)表示经过预处理后得到的图像和文本特征，$i = 1\sim N$。假设所提取的特征向量的维数相同，最简单的方法是采用直接叠加的方式将它们的特征向量加在一起，然后采用卷积运算来学习自适应权值，但是并没有改变原始的基网络[47]。在融合过程中，我们将得到的这两个分支中的嵌入图像、文本特征经过正则化处理分别表示为$S(X_i)$和$S(Y_i)$。融合后的视觉特征$f(X_i)$, 文本特征$g(Y_i)$可以由下式计算：

$$f\left(X_i\right)=W_I^{\,fuse}\odot S\left(Y_i\right)+b_I^{\,fuse} \tag{1}$$

$$g\left(Y_i\right)=W_T^{\,fuse}\odot S\left(Y_i\right)+b_T^{\,fuse} \tag{2}$$

其中，$W_I^{\,fuse}$和$W_T^{\,fuse}$是在多次训练中学习的融合权重，$b_I^{\,fuse}$和$b_T^{\,fuse}$为对应的图像、文本偏置向量，⊙代表内积运算。为了衡量融合的效果，一般用图像文本的匹配程度来说明，$f(X_i)$和$g(Y_i)$之间的匹配距离多用下面余弦距离公式来计算的：

$$d(f(X_i),g(Y_i))=1-\frac{f(X_i)\cdot g(Y_i)}{\|f(X_i)\|\cdot\|g(Y_i)\|} \tag{3}$$

匹配损失函数目的减少匹配对距离，增加非匹配对的距离，较小的匹配距离表示图像文本对存在某种关联[48]。所采用的直接融合只适应于维数相同的情况，但是针对神经网络维数不同的问题，通常处理方法是将某一模态的维数进行 *PCA* 降维处理至与另一模态相同的维数，然后再进行相同的融合操作[49]。仅仅通过上面计算余弦相似度来设置匹配函数是不够的，为了保持潜在空间中的相似性约束，通常需要在匹配损失上加约束函数。例如，文献[50]是基于一个类似于有效双向秩损失函数重新定义了匹配损失。为了利用更有代表性的非匹配对，该文在每一小批中选出了最具代表性的 K 类最不同的候选对象。直观地，这个损失函数是为了减小匹配对的距离和增加非匹配对的距离而设定的。损失函数的计算公式如下：

$$L_{mat}^{fuse}=\sum_{i=1}^{N}\sum_{k=1}^{K}\max\begin{bmatrix}0,d\left(f\left(X_i\right),g\left(Y_i\right)\right)\\-d\left(f\left(X_i\right),g\left(Y_{i,k}^{-}\right)\right)+m\end{bmatrix}+\alpha\max\begin{bmatrix}0,d\left(f\left(X_i\right),g\left(Y_i\right)\right)\\-d\left(f\left(X_{i,k}^{-}\right),g\left(Y_i\right)\right)+m\end{bmatrix}\tag{4}$$

m 为边缘参数，用来平衡两个三重因子，其中的$\mathrm{d}(\mathrm{f}(X_i),\mathrm{g}(Y_i)$表示匹配对的距离，$\mathrm{d}(\mathrm{f}(X_i),\mathrm{g}(Y_{i,k}^{-}))$，$\mathrm{d}(\mathrm{f}(X_{i,k}^{-}),\mathrm{g}(Y_i))$表示非匹配对的距离。将这一损失函数最小化将产生一个理想的潜在空间，其中匹配对的距离应小于任何不匹配对之间的距离。为了使用损失函数得到的结果直观的表示，使用了 *t-SNE* 算法可视化特征嵌入$f(X_i)$和$\mathrm{g}(Y_i)$，就可以得到图像文字特征的可视化表示。可视化结果表明：相匹配的图像文本在可视图中距离较近，不匹配的图像文本距离较远，该文所用的嵌入模型能够有效学习到图像文本的对齐表示，也即融合效果较好。

4.3 多模态对齐

在多模态学习中，除模态之间的融合外，模态对齐也是多模态学习的核心问题[51]。多模态的对齐负责对来自同一个实例的不同模态信息的子分支元素寻找对应关系。这个对应关系可以是时间维度的，例如电影画面、语音、字幕的自动对齐；对齐又可以是空间维度的，比如图片语义分割 ：尝试将图片的每个像素对应到某一种类型标签，实现视觉和词汇对齐。多模态对齐指的是分别处理多个单模信号，但在信号投影之前通过强制执行一定相似性约束来协调它们，即多模态表示中的协调表示，如下图 5 为模态对齐示意图。每种模式都有相应的投影函数，它们在一定相似性约束下互相对应。

目前针对多模态对齐，常见的两种分类为：隐式对齐和显式对齐[52]。隐式对齐一般是另一个任务的中间步骤，例如在基于文字的图像检索中，指单词和图像区域之间的对齐步骤，它确定了两种类型的隐式对齐模型；显示对齐是显式地将感兴趣的子模式之间的对齐[33]。它主要介绍如何实现不同子模式的对齐表示，以图像和声音两种模态作为研究对象，对于超过两种模态之间的对齐，采用两两对齐的方式以实现多模态对齐。

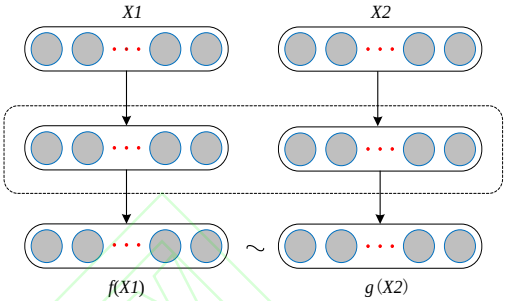

图 5　模态对齐结构示意图

在实际中，仅仅依靠相似度判别对齐实现的效果并不可靠，我们希望多模态的对齐表示既有一致性又有区分性，即判别对齐的准确度较高。目前有两种方法来解决这个问题：模型传递对齐和按等级对齐[53]。模型传递对齐是利用有区别的视觉模型来教学生模型一个有对齐的表示方法。以图像X_i和声音Y_i两种模态作为研究对象，例如X_i代表一个图像，Y_i代表图像对应的声音。用 $f_X(X_i)$和$f_Y(Y_i)$分别表示图像和声音模态的特征表示。假设$\mathrm{g}(X_i)$是某一特定模态类概率的教师模型，它可以估定特定模态的概率，由于各个模式是同步的，可以用另一种模式 $f_Y(Y_i)$来预测教师模型$\mathrm{g}(X_i)$的概率问题，使用 KL 散度作为损失函数计算公式如下：

$$\sum_{i}^{N}D_{KL}(g(X_i)\|f_Y(Y_i))\tag{5}$$

$$D_{KL(P\|Q)}=\sum_j P_j\log\frac{P_j}{Q_j}\tag{6}$$

这一目标本身将使对齐能够出现在$\mathrm{g}(X_i)$所预测的类别级别上。为了使内部表示出现对齐，需要限制网络上层跨模态的共享参数来实现，网络的上层参数在前期是特定于单个模态的，添加限制条件后上层参数将被各个模态共享，通过约束上层参数来转化为对齐表示[54]。为使对齐的区分效果更好，通常用按等级对齐方式的排序损失函数来获得有区分的对齐表示，该函数表示为：

$$\sum_{i}^{N}\sum_{j\neq i}\max\left\{0,\Delta-\psi(X_i,Y_i)+\psi(X_i,Y_j)\right\}\tag{7}$$

$$\psi(X,Y)=\cos(f_X(X),f_Y(Y)) \quad (8)$$

其中Δ代表边缘超参数，ψ是一个相似函数，j是迭代负例子。这一损失函数区别于前面仅仅靠余弦相似度判别对齐的好处在于，各自对齐的例子在表示空间中更加紧密的推到一起，达到一定的边缘设置参数。最后在三个基准实验上，在给定一个模式查询的情况下，在所有模式中都找到了相似的示例，验证了提出的对齐模型在视觉、声音和文本方面学到了更好的对齐[33]。

4.4 数据集

多模态深度学习具有极大的发展潜力，大量的研究在对现有的模型不断地进行改善和创新。除了寻求一切算法结构模型上的突破之外，不断更新完善数据集，提多模态深度学习模型运算速度，提高输出预测准确率高，对多模态学习的发展至关重要[55]。在本章列举常见的多模态任务相应的数据集，多模态学习区分单一模态在数据集上也有很大不同，下面介绍几种多模态常用的数据集。最初为了对会议室环境下说话人进行更好的研究，便于运用语音视频处理技术，需要大量的语音视频数据库。在这一领域收集的第一个里程碑数据集之一是 AMI 会议语料库，这是到目前为止信息量最多，功能最全面的音视频语料会议库，其中包含 100 多个小时的会议视频记录，每场会议由 4 到 5 个人组成，所有这些都经过了完整的转录和注释，以便人们更好的进行会议室环境下视频处理和语义分割等方面的研究[56]。另一个重要的数据集是信号语料库，主要研究说话者和听者之间的动态关系[57]。

这些数据集通常以人为中心的视觉理解，以及包括情感识别在内的变体，群体行为分析等[58]。例如：对于字母识别，avletters 是最常用的数据库之一，包含来自 10 个扬声器的录音，每个字母重复 3 次，分辨率为 376×288 像素和 25 帧[59]。后来又进行了改进，avletters2 解决了 avletters 的一些问题，例如低分辨率或扬声器数量有限[60]。具体来说，avletters2 增加了发声次数，每个扬声器重复 3 到 7 次和分辨率 1920×1080 像素和 50 帧。Pascal 数据集：它包含来自 20 个类别的 1000 幅图像(每类 50 幅)，其中一幅图像由五个不同的句子描述[61]。Flowers 数据集：包含 102 个类，共有 8189 幅图像。在训练阶段使用 2040 幅图像，其余 6149 幅图像用于测试[62]。CUB-Bird 数据集：它包含来自 200 个类别的 11，788 张鸟类图像,其中 5994 张图像用于培训，5794 张图像用于测试[63]。下表 3 为常用的多模态数据集。

表 3　多模态数据集

参考文献	年份	数据集	涉及模态	应用领域	数据集大小尺度	训练集测试集比
语音识别[15]	2015	IBM AV-ASR Large Vocabulary Studio Dataset	音频、视频	语音识别	大型	11:1
多模态情感分析[31]	2019	Getty、Twitter、Flickr	图像、文本	情感分析判别	大型	7:2
异构多媒体协同标注[23]	2018	Flickr、Youtube、Freesound websites、Google 3D Warehouse	图像、视频、音频、3D 模型	协同标注	大型	4:1
视觉文本匹配与分类[50]	2018	Pascal Sentence、MSCOCO、Flowers、CUB-Bird	图像、文本	匹配与分类	中型	7:1
深度对齐表示[33]	2017	Amazon Mechanical Turk	视觉、声音、文本	模态间对齐表示	中型	3:1

5 发展趋势与结论

关于目前的多模态深度学习，未来的发展趋势主要从以下几点说起：（a）探索如何应用神经网络研究多模态学习，还需要进一步研究形成一个以神经网络为基础的完善的理论体系结构，这取决于神经网络的理论体系的成熟发展；（b）与多模态相关的数据集也应该进一步完善，将直接决定深度学习模型的运算速度，输出预测准确率的高低，对多模态学习的发展也至关重要；（c）不同模态特征在融合过程中会受到噪声影响，使融合后信息不准确，并且在包含时序关系的多模态学习中，每种模态可能遭受噪声干扰的时刻也可能不同，因此在融合方式方面看是否还有更适合的方法；（d）现阶段的对齐方法显示对齐的数据信息量较少，且不同模态间信息甚至无法匹配使模型性能严重下降，在未来的工作中，还需设计同时进行度量学习和对齐的方法提高相模型的性能。

随着深度学习的快速发展，人们获取信息的方式的不断更新，由于信息数据的广泛性，数据库也不可能包含所需的全部信息，因此建立模态之间的相互联想能力格外重要，即使在数据信息不足，同样能够根据模态间的映射关系获取对事件的正确认知[64]。当然多模态应用很广，比较热门的研究方向用在自动驾驶汽车、多媒体应用和医疗领域等[65]。在这篇文章中，我们回顾了在深度多模式学习在视听语音识别、协同标注、匹配和分类以及对齐表示学习上的几个热门应用，对它们的具体实现过程作了简要概述，所提出的试听语音自动识别模型、统一空间映射模型、统一的多模式匹配和分类网络模型和跨模态对齐模型都有较好的实验效果。多模态学习是一个充满活力的多学科领域，具有日益重要和巨大的潜力。不可否认，将多种模式纳入学习问题会对网络结构、数据处理、目标函数设置等方面产生各种各样的影响，我们承认这在很大程度上是一个有很大挑战的领域，必然会出现许多新的创新，也期待着多模态学习领域这个方向更加蓬勃发展。

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，2017，11（13）：96-108.

[2] 邓智方，袁家政，刘宏哲，等. 基于浅三维稠密网的多模态手势识别算法[J]. 计算机工程与应用，2019，55（19）：166-172.

[3] 刘建伟，丁熙浩，罗雄麟. 多模态深度学习综述[J]. 计算机应用研究，2019，37（6）：56-75.

[4] Liu Huaping，Wang Feng，Zhang Xinyu，et al.Weaklypaired deep dictionary learning for cross-modal retrieval[J].Pattern Recognition Letters，2018，3（10）：66-73.

[5] 闫利，胡修兵，陈长军，等；多模态图像配准的梯度一致性算子[J]. 武汉大学学报(信息科学版),2013(8):969-972.

[6] 李盼龙；多模态核磁共振脑图像处理方法的研究及其应用[D]. 郑州大学，2019.

[7] Ma Lin，Chen Zhuo，Xu Long，et al.Multimodal deep learning for solar radio burst classification[J].Patter Recognition，2017，6（1）：573-582.

[8] Mohamed A，Dahl G，Hinton G，et al.Deep belief networks for phone recognition[C]//Nips Workshop on Deep Learning for Speech Recognition and Related Applications，2009，1（9）：25-39.

[9] Hinton G E，Osindero S，Teh Y W.A fast learning algorithm for deep belief nets[J].Neural Computation，2006，18（7）：1527-1554.

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

[11] Carletta J，Ashby S，Bourban S，et al.The AMI meeting corpus：A pre- announcement[C]//2nd International Workshop on Machine Leaming for Multimodal Interaction，2005，3869（5）：28-39.

[12] McKeown G，Valstar M F，Pantic M，et al.The SEMAINE corpus of emotionally coloured character interactions[C]// Proceedings of IEEE International Conference on Multimedia and Expo, 2010: 1079-1084.

[13] Hodosh M，Young P，Hockenmaier J，et al. Framing image description as a ranking task：Data，models and evaluation metrics[J].Neural Computation，2013，47（1）：853-899.

[14] 史秋萍. 基于 HMM 的视听语音识别系统[D]. 南京：河海大学，2011.

[15] Mroueh Y，Marcheret E，Goel V，et al. Deep multimodal learning for audio- visual speech recognition[J].IEEE Intelligent Systems，2015，978（7）：2130-2134.

[16] 梁冰，韩晶. 基于视听信息决策融合的自动语音识别方法[D]. 哈尔滨理工大学，2011.

[17] 邓佩，谭长庚.基于转移变量的图文融合微博情感分析[J]. 计算机应用研究，2018，21（7）：124-127.

[18] 韦航. 面向目标的中文微博情感分析研究[D]. 湖南大学，2018.

[19] Huang Feiran，Zhang Xiaoming，Zhao Zhonghua，et al. Image- text sentiment analysis via deep multimodal attentive fusion[J].Knowledge-Based Systems，2019，167（8）：26-37.

[20] Yi Bin，Yang Yang，Jie Zhou，et al.Adaptively attending

to visual attributes and linguistic knowledge for captioning[C]//Proceedings of the 2017 ACM on Multimedia Conference，2017：1345-1353.
[21] Yao Ting，Pan Yingwei，Li Yehao，et al.Boosting image captioning with attributes[C]//IEEE International Conference on Computer Vision，2017：4904-4912.
[22] Jin Zhiwei，Cao Juan，Guo Han，et al.Multimodal fusion with recurrent neural networks for rumor detection on microblogs[J].ACM Multimedia，2017，10（6）：795-816.
[23] Tian Feng，Wang Quge，Li Xin，et al.Heterogeneous multimedia cooperative annotation based on multimodal correlation learning[J].Journal of Visual Communication and Image Representation，2018，58（4）：533-544.
[24] Ling Zhenhua，Deng Li，Yu Dong. Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis[J]. IEEE Transactions on Audio Speech & Language Processing，2013，21（10）：2129-2139.
[25] 王景中，胡贝贝. 归一化算法在文字识别系统中的应用研究[J]. 计算机应用与软件，2011，28（3）：95-97.
[26] Niu Zhenxing，Zhou Mo，Wang Le，et al. Hierarchical multimodal LSTM for dense visual-semantic embedding[C]//IEEE International Conference on Computer Vision，2017，1（3）：1899-1907.
[27] Antol S，Agrawal A，Lu Jiasen，et al.VQA：visual question answering[C]//Proceedings of IEEE International Conference on Computer Vision.Piscataway，NJ：IEEE Press，2017：2425-2433.
[28] Li Huibin，Sun Jian，Xu Zongben，et al.Multimodal 2D+3D facial expression recognition with deep fusion convolutional neural network[J].IEEE Transactions on Multimedia，2017，19（12）：2816-2831.
[29] Xia Yingjie，Zhang Luming，Liu Zhenguang，et al.Weakly supervised multimodal kernel for categorizing aerial photographs[J].IEEE Transactions on Image Processing，2017，26（8）：3748-3758.
[30] Luong M T，Pham H，Manning C D.Effective approaches to attention-based neural machine translation[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing.Stroudsburg，PA：ACL Press，2015：1412-1421.
[31] Deng Cheng，Tang Xu，Yan Junchi，et al. Discriminative dictionary learning with common label alignment for cross-modal retrieval[J].IEEE Transactions on Multimedia，2016，18（2）：208-218.
[32] Niu Zhenxing，Zhou Mo，Wang Le，et al.Hierarchical multimodal LSTM for dense visual-semantic embedding[C]//IEEE International Conference on Computer Vision，2017：1899-1907.
[33] Aytar Y，Vondrick C.See，hear and read：Deep aligned representations[J].Antonio Torralba Massachusetts Institute of Technology，2017，5（1）：13-21.
[34] Song Zhichao，Ni Bingbing，Yan Yichao，et al.Deep crossmodality alignment for multi- shot person re- identification[J].ACM Multimedia 2017，61（8）：645-653.
[35] Wang Daixin，Cui Peng，Ou Mingdong，et al.Learning compact hash codes for multimodal representations using orthogonal deep structure[J].IEEE Transactions on Multimedia，2015，17（9）：1404-1416.
[36] Zhao Zhou，Lu Hanqing，Cai Deng，et al.Partial multimodal sparse coding via adaptive similarity structure regularization[J].ACM Multimedia，2016，19（3）：152-156.
[37] Li Zheng，Du Xiaobing，Ma Cuixia，et al.Interactive system for video summarization based on multimodal fusion[J].Journal of Beijing Institute of Technology，2019，28（1）：27-34.
[38] Morvant E，Habrard A，Ayache S.Majority vote of diverse classifiers for late fusion[C]//Proceedings of Structural，Syntactic，and Statistical Pattern Recognition.New York：Springer，2014：153-162.
[39] Chorowski J，Bahdanau D，Serdyuk D，et al.Attentionbased models for speech recognition[J].Future Generation Computer Systems，2015，10（4）：429-439.
[40] Jiang Xinyang，Wu Fei，Li Xi，et al.Deep compositional cross-modal learning to rank via local-global alignment[J].ACM Multimedia，2015，41（7）：69-78.
[41] 王保加，潘海为，谢晓芹，等. 基于多模态特征的医学图像聚类方法[J]. 计算机科学与探索，2018，12（3）：411-422.
[42] 童靖然，毛力，孙俊. 特征金字塔融合的多模态行人检测算法[J]. 计算机工程与应用，2019，55（19）：214-222.
[43] Seong T W，Ibrahim M Z.A review of audio- visual speech recognition[J].Journal of Telecommunication，Electronic and Computer Engineering，2018，10（1/4）：35-40.
[44] Li Zheng，Du Xiaobing，Ma Cuixia，et al.Interactive system for video summarization based on multimodal fusion[J].Journal of Beijing Institute of Technology，2019，28（1）：27-34.
[45] Baltrušaitis T，Ahuja C，Morency L P.Multimodal machine learning：a survey and taxonomy[J].IEEE Transactions on Pattern Analysis & Machine Intelligence，2017，41（2）：423-443.
[46] Ngiam J，Khosla1 A，Kim Mingyu，et al.Multimodal deep learning[C]//Proceedings of the 28th International Conference on Machine Learning，2015，10（6）：1-8.
[47] D'mello S K，Kory J.A review and meta- analysis of multimodal affect detection systems[J].ACM Computing Surveys，

2015，47（3）：1-36.
[48] Yin Zhong，Zhao Mengyuan，Wang Yongxiong，et al. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model[J].Computer Methods and Programs in Biomedicine，2017，140（4）：93-110.
[49] Graves A.Supervised sequence labelling with recurrent neural networks[M].New York：Springer，2012：1-131.
[50] Liu Yu，Liu Li，Guo Yanming，et al.Learning visual and textual representations for multimodal matching and classification[J].Pattern Recognition，2018，84（12）：51-67.
[51] Wang Limin，Qiao Yu，Tang Xiaoou.Action recognition with trajectory-pooled deep-convolutional descriptors[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.Piscataway，NJ：IEEE Press，2015：4305-4314.
[52] Srivastava N，Salakhutdinov R.Multimodal learning with deep Boltzmann machines[J].Journal of Machine Learning Research，2014，15（1）：2949-2980.
[53] Feng Fangxiang，Wang Xiaojie，Li Ruifan.Cross-modal retrieval with correspondence autoencoder[C]//Proceedings of ACM International Conference on Multimedia.New York，NY：ACM Press，2014：7-16.
[54] 刘宇鹏，马春光，张亚楠. 深度递归的层次化机器翻译模型[J]. 计算机学报，2017，40（4）：861-871.
[55] 梁斌，刘全，徐进，等. 基于多注意力卷积神经网络的特定目标情感分析[J]. 计算机研究与发展，2017，54（8）：1724-1735.
[56] Agrawal A，Lu Jiasen，Antol S，et al.VQA：visual question answering[J].International Journal of Computer Vision，2017，123（1）：4-31.
[57] Gao Lianli，Guo Zhao，Zhang Hanwang，et al.Video captioning with attention- based LSTM and semantic consistency[J].IEEE Transactions on Multimedia，2017，19（9）：2045-2055.
[58] 闫河，王鹏，董莺艳，等. 改进的卷积神经网络图片分类识别方法[J]. 计算机应用与软件，2018，35（12）：193-198.
[59] Cox S，Harvey Y.Lan，et al.The challenge of multi-speaker lip- reading[C]//International Conference on Auditory-Visual Speech Processing，2008：179-184.
[60] Rashtchian C，Young P，Hodosh M，et al.Collecting image annotations using amazon’s mechanical turk[C]//Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk，2010：139-147.
[61] Nilsback M E，Zisserman A.Automated flower classification over a large number of classes[C]//Indian Conference on Computer Vision，Graphics and Image Processing，2008：722-729.
[62] Wah C，Branson S，Welinder P，et al.The Caltech-UCSD birds-200-2011 dataset，CNS-TR[R].2011：21-33.
[63] 王景中，胡贝贝. 归一化算法在文字识别系统中的应用研究[J]. 计算机应用与软件，2011，28（3）：95-97.
[64] Wang Limin，Xiong Yuanjun，Wang Zhe，et al.Temporal segment networks：towards good practices for deep action recognition[J].ACM Transactions on Information Systems，2016，22（1）：20-36.
[65] Yu Dong，Seltzer M.Improved bottleneck features using pretrained deep neural networks[C]//Proceedings of International Speech Communication Association.Lisbon，Portugal：ISCA，2011：237-240.