计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2024/18(10)-2501-20
doi: 10.3778/j.issn.1673-9418.2403083

多模态数据融合研究综述

张虎成[1],李雷孝[1,2+],刘东江[1,2]
1. 内蒙古工业大学 数据科学与应用学院,呼和浩特 010080
2. 内蒙古自治区基于大数据的软件服务工程技术研究中心,呼和浩特 010080
+ 通信作者 E-mail: llxhappy@126.com

摘　要:尽管深度学习强大的学习能力已经在单一模态应用领域取得了优异成果,但研究发现单一模态的特征表示很难完整包含某个现象的完整信息。为了突破在单一模态上特征表示的阻碍,更大化利用多种模态所蕴含的价值,学者们开始提出利用多模态融合的方式去提高模型学习性能。多模态融合技术是让机器在文本、语音、图像和视频中利用模态之间的相关性和互补性融合成更好的特征表示,为模型训练提供基础。目前多模态融合的研究仍处在发展初期阶段,从近几年多模态融合的热门研究领域为出发点,阐述多模态融合方法和融合过程中的多模态对齐技术。重点分析多模态融合方法中的联合融合方法、协同融合方法、编码器融合方法和分裂融合方法在多模态融合中的应用情况与优缺点,阐述在融合过程中的多模态对齐的问题,包括显式对齐和隐式对齐以及应用情况与优缺点。阐述近几年多模态融合领域中热门数据集在不同领域的应用。阐述多模态融合所面临的挑战以及研究展望,以进一步推动多模态融合的发展与应用。

关键词:深度学习;多模态融合;模态对齐;多模态应用
文献标志码:A　　**中图分类号:**TP181

Survey of Multimodal Data Fusion Research

ZHANG Hucheng[1], LI Leixiao[1,2+], LIU Dongjiang[1,2]
1. College of Data Science and Application, Inner Mongolia University of Technology, Hohhot 010080, China
2. Inner Mongolia Autonomous Region Software Service Engineering Technology Research Center Based on Big Data, Hohhot 010080, China

Abstract: Although the powerful learning ability of deep learning has achieved excellent results in the field of single-modal applications, it has been found that the feature representation of a single modality is difficult to fully contain the complete information of a phenomenon. In order to break through the obstacles of feature representation on a single modality and make greater use of the value contained in multiple modalities, scholars have begun to propose the use of multimodal fusion to improve model learning performance. Multimodal fusion technology is to make the machine use

基金项目:国家自然科学基金(62362055);内蒙古自治区重点研发与成果转化计划项目(2022YFSJ0013,2023YFHH0052);内蒙古自治区高等学校青年科技英才支持计划项目(NJYT22084,NJYT24035);内蒙古自然科学基金(2023MS06008);内蒙古自治区直属高校科研项目(JY20220061,JY20222077,JY20230119,JY20230019);鄂尔多斯市重点研发计划项目(YF20232328)。
This work was supported by the National Natural Science Foundation of China (62362055), the Key Research and Development and Achievement Transformation Program of Inner Mongolia Autonomous Region (2022YFSJ0013, 2023YFHH0052), the Support Program for Young Scientific and Technological Talents in Higher Education Institutions of Inner Mongolia Autonomous Region (NJYT22084, NJYT24035), the Natural Science Foundation of Inner Mongolia (2023MS06008), the Research Projects of Universities Directly under Inner Mongolia Autonomous Region (JY20220061, JY20222077, JY20230119, JY20230019), and the Key Research and Development Program of Ordos (YF20232328).
收稿日期:2024-03-29　　**修回日期:**2024-06-13

the correlation and complementarity between modalities to fuse into a better feature representation in text, speech, image and video, which provides a basis for model training. At present, the research of multimodal fusion is still in the early stage of development. This paper starts from the hot research field of multimodal fusion in recent years, and expounds the multimodal fusion method and the multimodal alignment technology in the fusion process. Firstly, the application, advantages and disadvantages of joint fusion method, cooperative fusion method, encoder fusion method and split fusion method in multimodal fusion are analyzed. The problem of multimodal alignment in the fusion process is expounded, including explicit alignment and implicit alignment, as well as the application, advantages and disadvantages. Secondly, it expounds the application of popular datasets in multimodal fusion in different fields in recent years. Finally, the challenges and research prospects of multimodal fusion are expounded to further promote the development and application of multimodal fusion.

Key words: deep learning; multimodal fusion; modal alignment; multimodal applications

人工智能发展的灵感是模仿人类的多个感官从外界获取信息，例如视觉、听觉、嗅觉、触觉。由于人类生活在复杂信息相互交融的环境中，面临着大量、多样化的数据，获取信息更多以多维度的信息融合的方式，例如在自动驾驶领域，视觉传感器可以获取路况画面，激光雷达可以检测车辆前方物体的距离，声音传感器可以根据汽笛声做到听声定位，来弥补视觉盲区的道路信息。这些多模态数据蕴含着丰富的信息，对人们理解和处理现实世界的问题具有重要意义。随着文本图像的生成[1]、图文检索[2]、视觉问答[3]、机器人[4]和智能医疗[5]等领域发展，为了使机器能从多模态的信息中综合地提取信息，需要赋予机器理解、推理和学习的能力，因此需要将不同模态的信息进行融合，以获得更全面准确的分析和决策能力。

多模态数据是相同语义信息在不同的空间维度下的记录。而不同维度的数据具有不同的性质、结构和表征，这种异构性可能会影响模型的学习和应用。多模态信息融合的目标是减少这些异质性的差距，使用不同模态信息共同完成相同的任务。例如，对于同一信息的表达，既可以使用图像，也可以使用声音或文字。当某一模态数据丢失时，其他模态可以弥补这一缺失。这些数据之间的相互作用能够产生更完整的信息。它们以不同的方式相互作用产生更完整的信息。深度学习的发展是以数据为基础，多模态所蕴含的信息多于单一模态，优于单模态的学习为该领域的发展带来了全新的挑战[6]，例如在自动驾驶[7]、情感识别[8]和医疗诊断等领域。任何一个任务场景所传递的信息都不是单一的，而是以多种方式进行传递，不同的方式就会蕴含不完全一样的信息，利用好多种模态信息，挖掘更全面的信息，是多模态融合技术的研究目标，多模态融合适用于人工智能赋能的任何领域，是非常具有前景的领域。

本文按照联合融合方法、协同融合方法、编码器融合方法以及分裂融合方法四个层面对多模态融合进行分类，按照显式对齐和隐式对齐对多模态信息对齐进行综述，然后介绍公开数据集、面临的挑战以及未来展望。多模态数据融合研究框架如图1所示。

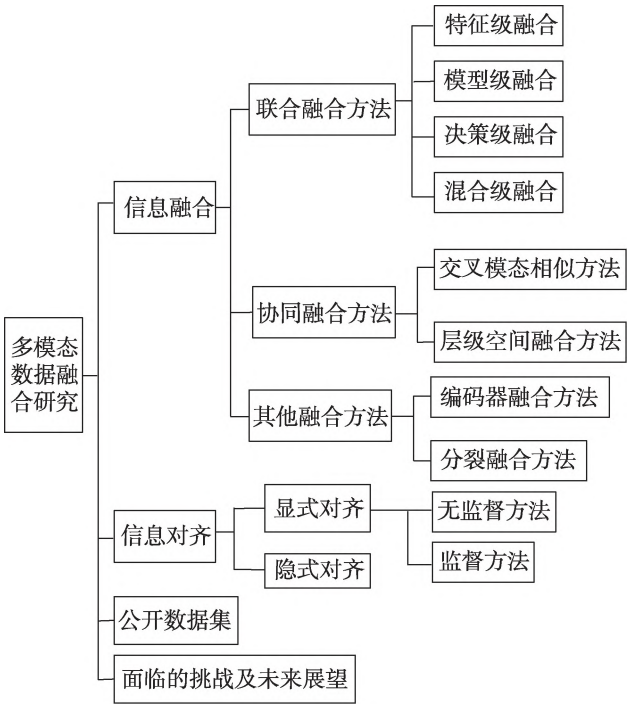

图1 多模态数据融合研究框架

Fig.1 Structure of multimodal data fusion research

1 多模态信息融合

多模态融合中表征的主要目的是对输入异质性数据进行统一的编码和表示，它可以理解为将原始的多模态数据转换为特定的数学表示形式或特征向

量的过程。

每个模态都有其独特的数据类型和表示方式，例如图像可以用像素值表示，文本可以用词向量表示，音频可以用波形数据表示。为了表征这些不同模态所蕴含的信息，找到一种有效的表示方式，使得不同模态的信息在表征空间中具有一致的语义或相关性，以便于最大化地利用不同模态信息去完成场景任务。这种表征通常会捕捉到数据的关键特征和信息，而丢弃冗余信息，从而提高模型的性能和泛化能力。

融合的方法按照其特点主要分为：联合融合方法、协同融合方法、编码器融合方法和分裂融合方法[9-10]。联合融合方法是通过将单模态的表示投射到一个共享的语义子空间中，该共享的语义子空间包含多模态的特征信息。协同融合方法旨在寻找多模态之间的关联关系并建立协调的语义子空间。编码器融合方法是将一种模态转化为另一种模态的表示形式。分裂融合方法旨在创建能够反映多模态结构的更大解耦表示集。融合的方式选择和设计直接影响到多模态融合的效果和性能。本综述将探讨多模态融合技术领域中主流的融合方法。

1.1 联合融合方法

联合融合方法是对每个模态的输入数据分别经过模态特定的编码器或特征提取器，得到单模态的表示，这些单模态表示被投影到一个共享的语义子空间。在这个共享语义子空间中，不同模态的特征可以进行联合融合、组合和进一步的处理[11]。联合模式相比较其他模式可以使各个模态的表示在共享的语义子空间中保持一致的语义信息，使得模态之间的关系更加紧密，能够减少特征维度的冗余，提取出对多模态任务贡献较大的重要特征，从而方便进行跨模态的特征融合和计算，联合融合方法示意图如图2所示。

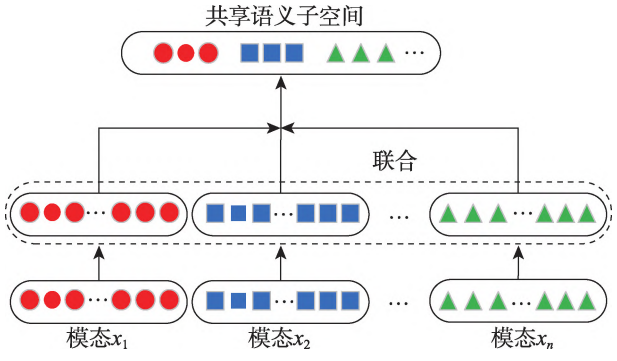

图2 联合融合方法

Fig.2 Joint fusion method

这种投影到共享语义子空间的操作可以发生在不同的融合阶段或融合时期，本文基于联合模式在不同阶段分为四种，特征级融合、模型级融合、决策级融合和混合级融合。在实际应用中，需要根据任务情况选择适合的融合模式方法，实验调整参数和结构，以获得最佳的多模态融合效果，对四种融合模式方法进行了比较如表1所示。

表1 四种融合模式方法性能比较

Table 1 Performance comparison of four fusion methods

方法	信息损失	融合难度	容错性	融合阶段
特征级融合	中	难	差	推理模型前
模型级融合	小	中	好	同时
决策级融合	大	中	中	子模型决策后
混合级融合	小	易	好	同时

1.1.1 特征级融合

特征级融合是在多模态数据输入到模型之前，将不同模态的原始数据或已从原始数据中提取的特征融合在一起，形成一个综合的表示来作为模型的输入。原始的数据蕴含不明显特征，因此原始数据和特征的融合均称为特征级融合。特征级融合方法示意图如图3所示。

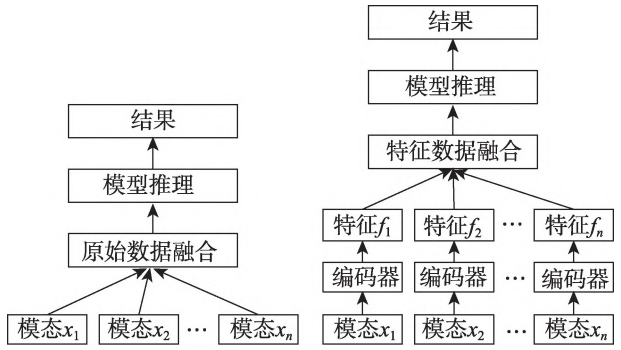

图3 特征级融合方法

Fig.3 Feature fusion methods

每个模态的数据首先经过各自的特征提取器或直接采用原始模态信息，例如图像可以使用卷积神经网络提取特征，文本可以使用词嵌入或文本卷积神经网络提取特征，音频可以使用声学特征提取方法。然后，将从不同模态的特征中得到的表示进行融合，特征级融合最常见的方法有拼接、加法、“乘”方法和双线性融合方法。

(1)拼接将多个特征向量进行简单拼接，得到一个更长的向量。简单直观，既不引入额外参数又保留多模态的原始信息。若模态数量较多时会导致维度

灾难,增加模型的复杂度和计算度。此外特征之间的融合不够充分,无权重系数也无法体现各态的重要性。

(2)加法将不同模态的特征进行线性加权求和,以融合多个模态的信息。假设有多个模态的特征向量,分别为 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$,其计算如式(1)所示:

$$\boldsymbol{y}=\boldsymbol{f}(\alpha_1\boldsymbol{x}_1+\alpha_2\boldsymbol{x}_2+\cdots+\alpha_n\boldsymbol{x}_n) \tag{1}$$

其中, $\boldsymbol{y}$ 是融合后的特征向量, $\alpha_1,\alpha_2,\cdots,\alpha_n$ 是权重系数,用于控制每个模态特征的重要程度,通过 $\boldsymbol{f}$ 将子模态特征转换为共享语义子空间。这些权重系数可以通过学习、手动设置或根据不同任务的先验知识来确定。加法融合引入权重系数可体现子模态重要程度,不足的是加法容易造成语义信息的丢失。

(3)“乘”融合方法将单模态特征向量相乘融合在统一的张量当中,使子模态语义信息充分融合。其计算公式如式(2)所示:

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

其中, $\boldsymbol{z}$ 表示融合张量后的结果输出, $\boldsymbol{v}$ 表示不同的模态变量, $\otimes$ 表示外积算子。“乘”融合可以弥补加法的语义信息丢失的不足。

(4)双线性融合中,首先将两个模态的特征进行张量外积,得到一个二维的双线性矩阵。通过展平操作将池化后的结果转换为一维向量。其计算公式如式(3)所示:

$$\boldsymbol{M}=\boldsymbol{f}(\boldsymbol{x}_1\otimes\boldsymbol{x}_2) \tag{3}$$

假设子模态特征向量分别为 $\boldsymbol{x}_1$ 、 $\boldsymbol{x}_2$,其中 $\otimes$ 表示张量外积算子, $\boldsymbol{f}$ 表示展平操作,能够较好地捕捉到不同模态之间的交互信息,并且保留了一定的空间结构和语义关联。

在特征级融合的思想上做特征选择,例如主成分分析、贝叶斯估计等。文献[12]在司机疲劳驾驶检测中,文中融合人的生理特征和行为特征共四种模态信息,使用特征级融合方法,直接输入到自编码器中进行训练权重,与单一模态相比该方法具有更高的准确度。文献[13]将四种多模态特征图按照通道进行拼接,拼接后使用卷积核压缩成与单模态相同的维度,实现异质性特征的互补,也提高了模型的准确率。

融合后的多模态共享语义信息可以继续输入到模型中用于任务的执行,特征级融合方法可以使模型直接利用多模态特征的组合信息,从而更好地捕捉到模态之间的关联和相互作用。可能面临模态间维度不匹配、信息失真等挑战。不同模态的数据具有不同的维度和尺度,因此在融合过程中需要进行适当的处理和归一化,避免出现中维度灾难,但是存在难以处理模态之间的时序性或局部关联的缺陷。特征级融合的实验对比如表2所示。

1.1.2 模型级融合

模型级融合(model-level fusion,MLF)是通过在模型级别上将不同模态的特征信息进行融合,实现跨模态的信息交互和整合。基于深度学习模型的融合方法应用范围更广且效果更好,也是目前研究者们首选的研究方法。常用方法包括早期的多核学习方法,该类方法目前适合小数据集的融合任务,然而经过深度学习的成熟发展,深度学习方法能够应对各种融合的场景。基于模型的融合方法是基于模型层面,但根据应用场景会与特征级或决策级没有明显的界线,特征级融合和决策级融合不属于模型级融合。模型级融合方法示意图如图4所示。

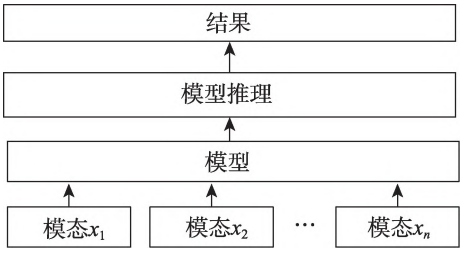

图4 模型级融合方法

Fig.4 Model-level fusion method

(1)多核学习方法

多核学习(multi-kernel learning,MKL)方法是在内核支持向量机方法的基础上改进的方法。内核支持向量机分类器在单模态特征空间中进行操作,在

表2 特征级融合的实验对比

Table 2 Experimental comparison of feature-level fusion

文献	任务	数据集	评价指标ACC/%
[12]	疲劳驾驶检测	沈阳地铁9号线延长线施工现场执行作业任务的司机中获取疲劳状态的多模态生理数据*	81.50
[14]	疲劳驾驶检测	2021年6月至7月在中国贵州省贵阳公共道路上收集的多模态驾驶数据集	99.21
[13]	农作物病害识别	PDR2018农作物病害数据集(番茄、苹果、樱桃、葡萄、柑橘、桃、草莓、辣椒、玉米、马铃薯)	98.44
[15]	农作物病害识别	番茄病害叶片数据(5 200张)和环境参数数据	98.90

不同的模态中选择不同的核函数时，需要为不同的核函数设置不同的权重参数，不同的模态数据集最主要的特性是异构性，数据集的异构性会让模型难以达到理想效果。在应用中，为了弥补上述的不足，MKL应运而生。其目的是学习一组预定义的基本核的线性或非线性组合，多核映射作用下，高维空间成为多个特征空间组合而成的组合语义空间。组合语义空间充分利用基本核的不同特征映射能力，组套索正则化器的使用可以确定每个基核的最优权值，以便发挥每个基本核的最大能力。由于核可以看作各数据点之间的相似函数，从而实现特征选择融合，选择该方法能更好地融合异构数据且使用灵活[16]。文献[17]采用（multi-kernel support vector machine，MK-SVM）从不同模态中选择特征进行融合，在阿尔茨海默病神经成像计划数据集的脑成像遗传数据上进行实验探索并进行最终诊断和预测。文献[18]将MKL方法对声学、语义和艺术家的社会观三方面进行音乐艺术家相似性排序，学习相似的空间项目来产生相似的空间，结果显示以最优方式将所有特征信息组合到一个共享的嵌入空间中。图5为多核学习方法示意图。

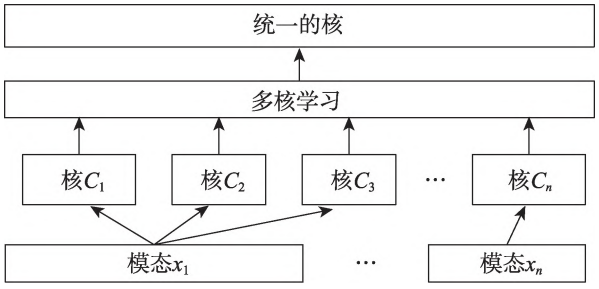

图5 多核学习方法

Fig.5 Multi-kernel learning methods

MKL的基本核函数选择灵活外，损失函数为凸函数也是优点之一，便可以使用标准优化包和全局最优解进行模型训练，其计算速度和模型性能可以得到大幅度提升。MKL的不足是在测试期间需要依赖训练数据，这意味着对于新的测试样本，必须重新计算和调整核函数的权重，这可能会增加计算的复杂性和时间开销，MKL在训练阶段需要占用大量的内存资源，这可能会限制其在资源受限环境中的应用。

（2）基于神经网络的融合

基于神经网络的融合方法是现在主流的研究方向，其可以融合不同模态的异构数据，该模型还可以融合不同的图像，例如自动驾驶领域激光雷达图像和摄像头捕捉的视觉的图像，在医疗领域的不同设备采集的不同影像图像等。基于神经网络的融合具有效率高、学习能力强的优势已经在很多领域得到了广泛的应用。本文主要阐述模态级融合的生成对抗网络和基于注意力机制的融合方法。

①生成对抗网络是深度学习领域中无监督学习数据的方法，受到多模态融合技术的启发，基于生成对抗网络逐渐有了发展。文献[19]基于多特征融合的生成对抗网络用于水下的图像增强，该算法利用多特征的信息，显著提高水下的图像质量。文献[20]提出基于去噪扩散模型（denoising diffusion probabilistic model，DDPM）的多模态图像融合方法，该模型由无条件生成模块和条件似然校正模块组成。融合图像的采样仅通过预训练的DDPM实现，不需要进行微调，能够生成融合图像包含每个模态的互补信息，在红外光与可见光的融合和医学图像融合实验中得到证实。

文献[21]使用基于生成对抗网络的子网获得模态级原型融合其他模态特征用于手势识别，模态信息的流通有助于利用它们的互补性，并且可以优于最先进的方法。文献[22]提出跨模态对比生成对抗网络应用于文本合成图像，使用基于注意力机制方法的生成器，强制执行强文本-图像对应，以及一个对比判别器，作为对比学习的评论家和特征编码器，可以生成更高质量的图像，更好地匹配输入描述。

②注意力机制在自然语言处理领域和图像处理领域的成功表现，得益于具有全局感受野的能力。注意力机制在多模态特征融合任务中具有明显的优势，它可以从原始输入中选择显著的特征，并帮助处理存在噪声、语义分歧和语义重复等问题。通过注意力机制，模型可以根据各模态之间的关系动态地评估它们的重要性，并提取出模态之间的互补信息。这些信息被整合到一个单一向量表示中，从而缓解了语义歧义的问题。换句话说，注意力机制可以帮助模型更加准确地理解和融合多模态特征，提供更准确和全面的表示。

文献[23]在自动驾驶领域为了提高三维目标检测的性能，使用特征级融合图像二维数据和激光雷达三维数据，简单融合后使用坐标注意力机制融合的方法挖掘深层次的特征信息。该方法保证了深度特征映射也包含更丰富的语义信息，而浅层特征融合保留了更完整的几何细节。该方法使用SimAM无参数注意机制来评估网络中每个神经元的重要性。

这种机制通过定义能量函数的形式来实现分化，其中具有更高能量函数的神经元由于其重要性的增加而被赋予更大的权重。相比之下，那些具有较低能量函数的函数可以赋予较低的权重。

文献[24]在抑郁症的检测领域中，医疗设备所采集的文本、音频和视觉对患者的诊断有重要的参考价值，在三个模态融合前提取每个模态数据的预先设计的特征和高级语义，然后用循环神经网络和一维卷积神经网络分别设计单模态模型对特征进行编码。然后使用改进的Transformer网络将同一模态的预先设计的特征编码和高级语义编码相融合，以得到各模态特征的更好表示。进一步使用模态内注意机制融合不同模式的特征，将模态内的融合作为主要特征，将预先设计的特征作为辅助特征。最后利用融合结果完成抑郁症分类，相比较于简单拼接多模态数据可以解决数据冗余的不足。

文献[1]提出了多尺度通道注意力模块(multi-scale channel attention module，MS-CAM)以解决融合不同尺度特征困难的问题。MS-CAM的关键特点在于全局平均池化和点卷积来获取全局特征和局部特征的通道注意力权重。根据不同的网络场景提出了基于MS-CAM的通用注意力特征融合方法，使得网络能够进行软选择或加权平均来融合不同尺度的输入特征，以实现更好的融合性能。

Transformer是多头注意力机制的应用，在计算机视觉与自然语言处理领域非常成熟，并且对文本与图像的编码过程具有相同的原理。为图像-语言的预训练模型的双模态融合预训练提供理论的基础。文献[25]提出基于Transformer模型融合文本与图像，该方法使用各自特征编码器把文本与视觉的Embedding分别输入Transformer得到包含位置信息和特征编码信息，将两个模态的输出进行线性映射，然后输出结果拼接，加上一个位置编码信息与原Transformer的结构一致，便可以进行图像文本的融合训练。

文献[26]提出一个通用有效的预训练策略，使用轻量化的转换器弥补通道之间的差距。该转化器分为两个阶段预训练，第一个阶段从固定的图像编码器中引导视觉语言表示学习，第二阶段是从固定的语言模型中引导视觉到语言的生成学习，从而完成语言文本与视觉图像的信息融合。该方法性能和计算效率得到显著提升。

文献[27]提高图像理解和语言生成任务的性能，文中提出PromptFuse和BlindPrompt方法。PromptFuse方法通过在预训练模型中添加可学习的提示向量来引导模型生成与图像相关的语言描述。BlindPrompt方法则是在预训练模型中添加固定的提示向量，不需要进行微调。这两种方法都能够在少量参数的情况下实现图像和语言的融合，提高模型的性能。

注意力机制可根据不同任务和情境动态地调整不同模态的注意力权重，使模型能够适应不同的输入和输出需求。通过注意力权重的分配，注意力机制能够突出显示模型在决策过程中的重要输入，提供对模型决策的可解释性。多模态数据中的不同模态通常包含不同的信息。可以根据任务需要自适应地对不同模态进行加权融合，以获得更具信息量的特征表示。注意力机制可以减轻来自其他模态的噪声的影响，提高模型在处理多模态输入中的鲁棒性。注意力机制在多模态融合中具有灵活性、可解释性、强大的特征表达和抗噪性等优势，从而提升了模型在多模态任务中的性能表现。

③双线性池化(bilinear pooling，BP)是一种通过计算特征向量的外积来进行多模态特征融合的方法。它可以创建一个融合表示空间，在这个空间中充分利用向量元素之间的交互作用。双线性池化也被称为二阶池化，与简单的向量组合操作不同，它在同一位置的两个特征双线性相乘，得到 n^2 维矩阵 $\boldsymbol{b}$ 表征向量，对这 n^2 向量进行归一化操作和L2归一化操作，便得到双线性池化后的融合特征。这使得双线性池化方法更具表现力。在实际应用中双线性池化方法常与注意力机制相结合，将融合的双模态表示作为注意力模型的输入特征，文献[28]提出了一种利用文本和图像信息进行特征融合的多模态假新闻检测框架。该框架使用文本和视觉的单模态特征提取器提取特征，并试图最大限度地提高文本和图像之间的相关性，之后使用多模态分解双线性池进行特征融合，从而获得高效的多模态共享表示。该方法解决文本特征和视觉特征融合性能不足的问题。文献[29]实现RGBT跟踪的关键是融合RGB图像和热模态不同层次的抽象信息。现有的文献关于RGBT跟踪算法要么专注于最后一层的语义信息，要么使用简单的操作从每个模态聚合层次深度特征，这限制了多模态跟踪的能力。该文在实现分层特征融合之前，利用通道注意机制对所有卷积层特征实现特征通道的自适应标定，通过叉乘对任意两层进行双线性池化操作，这是一种二阶计算，有效地聚合了目标的深层语义和浅层纹理信息，设计质量感知融合模块，以自适应方式聚合不同模式和不同层相互作用的双线性池化特征。该方法相较于现有方法，提

升效果显著。

模型级融合可以减少模态间的不匹配问题和维度差异问题，提高模型的鲁棒性和泛化能力。但需要更多的计算资源和存储空间，每个模态都需要单独的网络结构和参数，意味着调参和优化变得更难。同时，不同模型的输出可能存在尺度、格式上的不匹配，需要额外的预处理或转换步骤来保证融合的有效性。最终的性能可能高度依赖于各个单独模型的性能，例如其中一个模型性能不佳，可能会拖累整体性能。模型级融合需要设计合适的融合策略和网络架构，以有效地整合不同模态的信息共同发挥价值。模型级融合的实验对比如表3所示。

表3 模型级融合的实验对比

Table 3 Experimental comparison of model-level fusion

文献	任务	数据集		评价指标	
[17]	阿尔茨海默病诊断	ADNI脑成像遗传数据集		ACC/%	96.48
[30]	阿尔茨海默病诊断	ADNI脑成像遗传数据集		ACC/%	95.89
[19]	水下图像增强	EUVP数据集		水下图像质量度量(UIQM)	更优
				水下彩色图像质量评价(UCIQE)	优
[31]	水下图像增强	UIEB数据集、UFO-120数据集		UIQM	更优
				UCIQE	更优
[20]	图像生成融合	ImageNet数据集预训练	在TNO、RoadScene、MSRS和M3FD数据集进行红外光与可见光实验	熵、标准差、互信息、视觉信息保真度、QAB/F和结构相似指数测度	提升
			在MRI-CT、MRI-PET和MRI-SPECT图像对进行医学图像融合实验		提升
[32]	图像生成融合	TNO、Newdata_gary、Newdata_color和RoadScene数据集上进行红外光与可见光实验		加权融合质量指数 Q_w、结构相似性 Q_{SSIM}、Chen-Varsheny Q_{CV}、加入融合图像的噪声的比例 N_{abf}	显著提升
		MRI图像和不同病理状态的多种模式脑图谱数据进行医学图像融合实验			
[21]	手势识别	Chalearn IsoGD数据集		ACC/%	71.52
		EgoGesture数据集		ACC/%	98.39
		THU-READ数据集		ACC/%	86.67
[33]	手势识别	RWTH-PHOENIX-Weather 2014数据集		ACC/%	73.9
		American Sign Language Lexicon Video 数据集		ACC/%	97.0
[22]	文本合成图像	MS-COCO数据集		FID	9.33
		LN-COCO数据集		FID	14.12
[34]	文本合成图像	CUB bird数据集		FID	13.32
		COCO数据集		FID	15.83
[23]	三维目标检测	DAIR-V2X数据集		对行人和自行车检测的AP/%	提升≈7
				在IOU=0.7时车辆检测的AP/%	≈80
[24]	抑郁症诊断	中国多模态抑郁症语料库		F1-score/%	95
		EATD-Corpus		F1-score/%	50
[35]	抑郁症诊断	AVEC2013数据集		平均绝对误差(MAE)	6.60
		AVEC2014数据集			7.05
[28]	假新闻检测	Twitter数据集		ACC/%	88.3
		微博数据集(2012年5月至2016年6月)		ACC/%	83.2
[36]	假新闻检测	Weibo16数据集		ACC/%	94.2
		Weibo23数据集			84.5
		Wechat18数据集			88.1
[29]	RGBT跟踪	RGBT234数据集		精度率(PR)/%	88.5
				成功率(SR)/%	71.6
[37]	RGBT跟踪	GTOT数据集		精度率(PR)/%	90.3
				成功率(SR)/%	77.2
		RGBT234数据集		精度率(PR)/%	70.1
				成功率(SR)/%	69.9

1.1.3 决策级融合

决策级融合(decision-level fusion,DLF)将每个模态的独立决策结果进行数学公式规定或赋予不同结果不同的权重来得出最终的决策结果。常见的决策级融合策略包括投票法、加权平均法和多数投票法等。决策级融合方法示意图如图6所示。

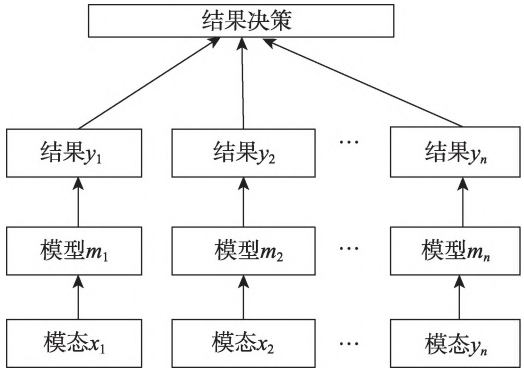

图6 决策级融合方法

Fig.6 Decision-level fusion methods

投票法通常将多个模态的独立决策结果进行投票统计,选择获得最高票数的类别或结果作为最终的决策。文献[38]在文本-情感分析模型中聚合三种模型的决策,得分最高的选票被认为是对拟议框架的预测情绪。然后本文为每个样本计算了一个情绪得分。这些情感得分可与分类器所作的决策进行决策融合,从而实现决策融合机制。加权平均法将不同模态的决策结果按照权重进行加权平均,得到一个综合的决策结果。文献[39]将由支持向量机获取的三类特征的后验概率进行加权融合,并将权重求解过程转化为粒子群优化算法的全局寻优。这种融合方法能够更好地提高决策的准确性和性能。文献[40]提出了一种基于模型可靠性的决策级融合策略,为了提高可见光和红外图像决策级融合目标检测算法的性能。多数投票法则是根据各个模态的决策结果中出现的频率最高的类别或结果进行决策。相较于之前的早期融合方式,这种融合方式具有处理简单数据异步性的能力,其优势在于允许使用最适合其中单模态的提取特征的方法。文献[41]提出了一种称为Bi-Bimodal Fusion Network的融合方案,是将多个模态源数据结合起来,对其标签进行预测。文献[42]提出了一个名为两阶段多任务情感分析的多模态框架。它采用两阶段训练策略来充分利用预训练模型和一种新的多任务学习策略来研究每个表征的分类能力。

决策级融合方法忽视了模态之间的相互作用和关联性,也难以利用模态之间的互补性,该方法需要为每一个模态训练分类器,学习过程变得耗时且复杂。换句话说,每个模态单独地训练权重,最后整体权衡各子模态的结果。后期融合的处理与特征无关,需要多个网络模型进行训练,能够很好地适应模态缺失问题,有更大的容错性。决策级融合的实验对比如表4所示。

表4 决策级融合的实验对比

Table 4 Experimental comparison of decision-level fusion

文献	任务	数据集	评价指标	
[38]	情感分类	Assamese数据集*	ACC/%	95.1
[40]	图像融合	KAIST多光谱行人数据集	白天漏检率/%	降低8.16
			晚上漏检率/%	降低9.85
[41]	情感分析	CMU-MOSI、CMU-MOSEI和UR-FUNNY数据集	(平均绝对误差、7级精确度、ACC2和F1-score)/%	提升≈1
[42]	情感分析	CMU-MOSI数据集	ACC2/%	87.0
		CMU-MOSEI数据集	ACC2/%	85.6

1.1.4 混合级融合

混合融合方法综合特征级融合、模型级融合和决策级融合方法三种融合方式的优点,在不降低性能的同时,也可以根据应用场景的融合难易程度选择合适的组合。混合级融合方法示意图如图7所示。

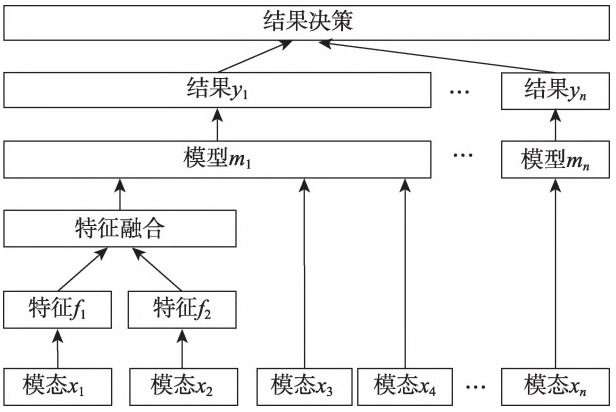

图7 混合级融合方法

Fig.7 Mixed-level fusion methods

混合级融合能够综合多个模态的信息,全面地利用不同模态的特点和信息,从而提高了融合结果的综合性和鲁棒性。在应用中,混合级融合方法具有一定的灵活性,能够根据不同的应用场景和数据特点灵活选择融合的方法和策略,提高了系统的适

用性和灵活性。同时也面临一些挑战：融合过程相对复杂，可能需要更多的计算资源和时间成本；参数选择和调优相对困难，需要花费较多的精力和时间来确定最优的参数组合；不同的混合级融合方法对于不同的数据和任务效果并不确定，需要在具体问题中进行实验和验证，方能确定最适合的融合策略。混合级融合方法在充分利用不同模态信息的同时，需要在实际应用中综合考虑其优缺点，并进行合理的选择和平衡。

1.2 协同融合方法

协同融合方法是使用约束条件作用在各个单模态中，使其模态之间相互协同。协同的目标是确保不同模态之间的信息相互补充、相互支持[11]。不同模态的特征有异质性的特性，其包含的信息也是不平等的，学习分离表征有益于保持模态特有的排他和有用的特征[43]，并且在整体融合结果中发挥协同作用。由于协调表示学习保留了原始模式的信息，且其优化目标是不同模式之间的合作关系，它适用于仅以一种模式为输入的应用。而联合表示学习最终只能得到统一的表示。其最终优化目标是模型预测性能，适用于多模态输入。目前的在多模态融合领域中，协同融合方法主要分为交叉模态相似方法和层级空间融合方法。协同融合方法示意图如图8所示。

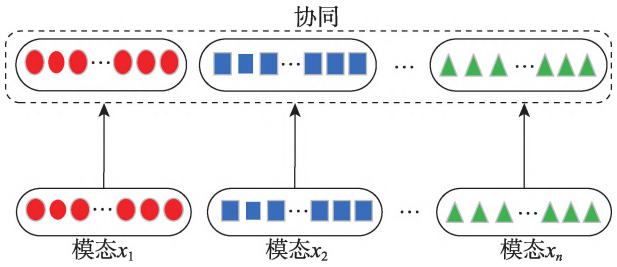

图8 协同融合方法

Fig.8 Collaborative fusion methods

1.2.1 交叉模态相似方法

交叉模态相似性方法用于比较不同模态数据之间的相似性。旨在通过计算子模态之间相似性来量化它们之间的关系[44]。

期望与同一语义或对象相关的模态间相似度距离尽可能小，而与不同语义相关的模态间相似度距离尽可能大。例如单词“Dog”所表示的含义要和一只狗的图像表示接近，而单词“Dog”和一辆车的图像要远离[45]。当度量两个文本文档之间的相似度时，余弦相似度比基于距离的度量更有用。由于两个文档之间有更多相同的单词，两个文档不共享大多数单词的可能性非常高，词频用矢量表示。余弦相似度则表示为两个向量之间的夹角[46-47]。其计算公式如式(4)所示：

$$S(\boldsymbol{x},\boldsymbol{y})=\cos\theta=\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\cdot\|\boldsymbol{y}\|} \tag{4}$$

交叉模态相似方法可以避免对高维特征向量的处理。在多相似度融合中，融合的相似度值可以通过式(5)融合规则计算：

$$S_{\text{fusion}}(\boldsymbol{x},\boldsymbol{y})=F_{c=1}^{k}(S_C(\boldsymbol{x},\boldsymbol{y})) \tag{5}$$

其中，S_C，$c=1,2,\cdots,k$。k为特征向量$\boldsymbol{x}$和$\boldsymbol{y}$之间的不同相似度度量，F为融合规则，包括MIN、MAX和加权AVG融合。MIN和MAX规则表示相似性的极端值分配，但对“噪声”存在高敏感度，加权AVG规则对噪声点具有更好的稳定性。MIN和MAX规则可能比加权AVG规则更具区别性[48]。常见的交叉模态相似性方法包括损失函数[49]、相关系数、基于信息熵的相似度、基于统计距离的相似度。

交叉模态相似性方法可以更深层次地理解和解释数据，因为它允许模型从一个模态中学到的知识应用到另一个模态上，单一模态可能无法提供足够的信息来完成特定任务。交叉模态相似方法通过整合不同模态的信息，可以克服这一限制。不足是同模态之间存在数量或质量上的不平衡会影响到模型学习过程中各模态信息的有效整合。

1.2.2 层级空间融合方法

层级空间融合将不同模态的数据在不同层次上进行融合，以捕捉它们之间的相关性和互补性[50]。文献[51]在层次空间中将图像和文本视为两层并执行顺序嵌入的目的是捕捉文本和图像嵌入的部分顺序，以在协调空间中执行层次结构。顺序嵌入可以在一个单一的模型中学习到整个特征语义。文献[52]也提出了一个使用表示图的类似模型，其中表示图用于引入这样的层次结构。文献[53]提出层级特征融合方法应用于多模态情感分析，该策略以一层的方式进行，先融合两种模态，其后融合所有模态。该方法相较于特征级融合能够学习双峰和三峰相关性，并使用深度神经网络进行数据融合。文献[54]提出了一种分层特征融合的连接结构模型，应用于检测路面裂纹，主干网络中间层特征图与高层特征图相结合，从而提高具有更多语义信息的高级特征映射的分辨率，作为主干网络检测裂纹。注意层融合低层特征图和高层特征图，可使高层预测图具有更好的定位和更清晰的裂纹边界，注意力层用于恢复裂纹的边缘信息。文献[55]提出新的分层多模态融合

方法，用于癌症生存预测。该方法采用多种融合策略，将融合问题分解为不同的层次，每个层次从低层次向高层次逐步整合和传递信息，从而使融合过程更加特殊化，多模态表达更加丰富，该方法相较于现有方法显著降低计算复杂度。文献[56]提出了一种基于层次多模态融合的危机事件摘要生成模型，文中将文本特征向量和视觉特征向量生成对应的文本上下文向量和视觉上下文向量，将每个视觉上下文向量与文本上下文向量层次融合，保持图像信息的独立性。然后生成多模态上下文向量，在一定程度上提高了危机事件多模态输出的质量，并提高了最终输出的质量。文献[57]为了提高多模态特征的融合和分割精度，提出了一种分层多模态融合的多任务感知网络来学习RGB-T城市场景。并开发层次多模态融合模块来增强特征融合，并构建了高级语义模块来提取语义信息，以便在不同抽象级别上与粗特征进行合并，利用多级融合模块，利用低、中、高级融合来提高分割精度。该方法相较于传统的融合方法具有效率高、融合充分的优点。

层次空间融合的优势在于能够充分利用不同模态数据的互补信息，提高模型的表达能力和性能。同时，不同层次的数据可能面临规模或时间序列上的不一致问题，需要额外的预处理步骤来确保各层次信息的有效整合，如果在高维特征空间进行过于复杂的数据融合，可能导致模型过拟合。

1.3 其他融合方法

除上述的融合方法外，还有一些不是主流的融合方法，如深度学习发展早期出现的编码器融合方法和近几年出现的分裂融合方法。编码器方法已经不能独当一面地在融合中被使用，分裂融合方法也还未发展成熟。

1.3.1 编码器融合方法

编码器融合方法用于将一个模态特征映射到另一个模态的特征的转化任务中。它由编码器和解码器两个部分组成[58]。该方法中编码器将输入的初始模态转换为一个向量表示。这个向量包含了初始模态的有用特征数据。解码器使用以前的向量作为输入，生成一个新的模态，这种架构的目的是将不同模态之间的信息转换和融合，从而实现模态之间的相互转换和互补。编码器融合方法示意图如图9所示。

文献[59]自适应解码器的多模态对话系统，该系统首先使用多模态上下文编码器嵌入历史话语，根据编码器将用户归为不用的类别和意图，然后在统一的解码器中编码各种形式的领域知识以响应用户的意图。文献[60]设计了多模态融合方法去生成逼真的人脸动作模型，该方法使用编码器融合方法去融合声音和眼部信息。

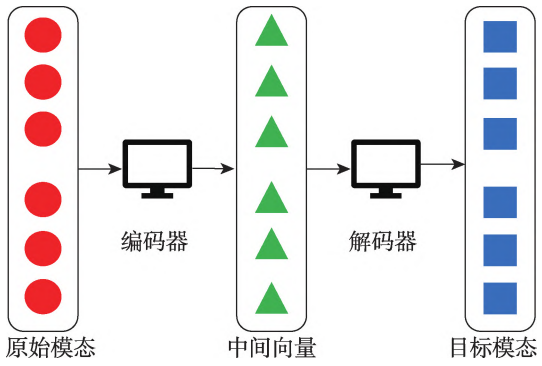

图9 编码器融合方法

Fig.9 Encoder fusion method

编码器融合方法的优点是可以生成新的模态输出，但缺点是每个编码器只能处理一种模态，可能增加模型复杂性，并且对于配对数据稀缺的模态存在挑战。在应用中需要权衡这些因素，并选择适合具体任务和数据情况的方法。

1.3.2 分裂融合方法

分裂融合方法旨在创建一个新的分离表示集，该表示集相较于输入表示集会更大，反映出内部多模态结构的知识[50]，如数据聚类、独立变异因素或模态特定信息。与联合和协调表征相比，分裂融合方法具有精细解释和细粒度可控性。根据解耦因子的粒度，方法可以分为模态级分裂和细粒度分裂。分裂融合方法示意图如图10所示。

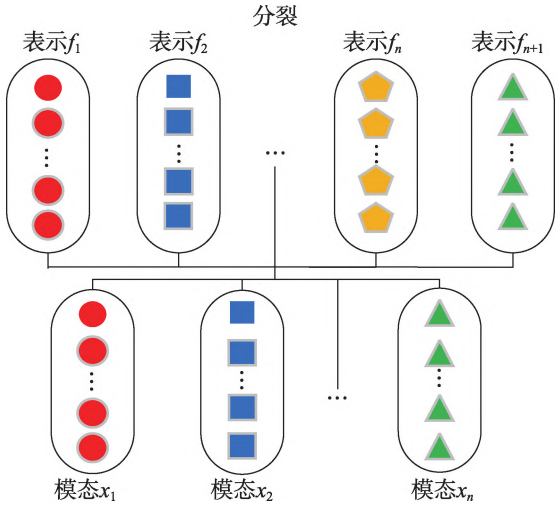

图10 分裂融合方法

Fig.10 Fission fusion methods

模态水平分裂的目的主要是在每个模态中分解出特定于模态的信息，在两个模态中分解出冗余的多模态信息[61]。然后将它们融合在一起以获取更全面和准确的信息。这种方法可以提高多模态数据融合的效果，因为它允许每种模态的特点得到更好的发挥，并且可以采用专门针对每种模态的处理技术。文献[62]提出多模态变分自编码器中的模态级分裂去学习不同模式集的生成模型。模态级分裂的目的是最大化每种模态的信息利用，从而提高整体多模态融合系统的性能。

细粒度分裂是指将每个模态的信息进一步细分为更小的单元或特征。这种分裂可以在不同级别上进行，例如在图像中可以将其分解为区域或局部特征，而在文本中可以将其分解为单词或短语。通过细粒度分裂，可以从多个角度捕捉到每个模态中的更多细节和特征，从而更详细地描述每种模态的内容。文献[63]使用多模态自我监督学习在没有监督标签的情况下训练大型网络用于视频聚类。

细粒度分裂的目的是为了提供更丰富和准确的特征，以增强多模态融合系统的性能和效果，细粒度分裂可以应用于不同的多模态任务。

2 多模态信息对齐

多模态融合中，模态之间的信息对齐是重要的核心问题之一[64]。模态对齐问题寻找不同模态信息中来自同一实例的子分支元素之间的对应关系，可以促使学习到的多模态表示更加精确。对齐的关系可以是时间维度的，例如自动对齐电影画面、语音和字幕；也可以是空间维度的，比如在图像语义分割任务中，试图将每个像素对应到某一种类型的标签，完成视觉和词汇的对齐[65]。多模态对齐是在对各单模态数据进行特征提取之前，通过强制执行一定的相似性约束来协调它们，从而得到协调的多模态表示。每个模态都有相对应的映射函数，按照有无对应的标签主要分为显式对齐和隐式对齐[66]。

2.1 显式对齐

显式对齐主要处理模态之间子组件的对齐问题。该方法通过明确的算法或标签进行直接关联和对齐不同的模态信息，其目的是提供模态之间的对齐关系，不需要额外地设计模型让信息对齐，提高了信息整合的效率和准确率。其中，显示对齐包含无监督对齐方法和监督对齐方法。

2.1.1 无监督方法

无监督方法在子组件对齐任务中不依赖人工标注的标签。文献[67]提出动态时间扭曲的无监督对齐方法用于解决多视图时间序列。文献[68]提出通过相同物体的外貌特点定义视觉场景和文本之间的相似性来解决电视节目和情节内容的对齐问题。在情感识别领域，词级强制对齐是将各种模态信号分割成相应的词序列，然后进行模态间交互和融合操作。这种对齐只关注基于词的局部融合，忽略了序列关系之间的交叉词的影响。例如文献[69]提出了一种情感识别模型，该模型在词层面提取不同的模态特征，并采用强制对齐来实现模态间的时间依赖性交互。基于顺序关系的对齐全局检测词与帧之间的关系，实现多模态融合和交互。文献[70]设计了一种时间对齐均值-最大值汇集机制，以探索隐藏在模态细粒度特征空间中的情感相关线索。文献[71]引入了一种时间语义交互网络，该网络具有细粒度的时间对齐机制，用来模拟框架和单词之间的时间依赖性。由于无监督对齐方法不需要依赖标注数据，可节约数据标注的成本，但对不同模态之间的数据规范性要求较高，需要保持时序一致性，并且时序上不能存在较大的波动和不连续性，否则对齐性能会显著下降。

文献[72]提出了一种解耦特征对齐与融合的框架。这个框架分为两个阶段：在第一阶段，进行无监督表示学习，并使用模态自适应模块来对齐来自各种模态的特征；第二阶段，自注意机制融合模块利用监督学习将医学图像特征与临床数据相结合。该框架成功应用于医疗领域疾病的预测。文献[73]提出了一个完全无监督的网络对齐框架。基于边缘轨道的高阶拓扑一致性并将其融合到图卷积网络的信息聚合过程中，将对齐一致性转化为节点嵌入的相似性，该方法对结构噪声具有鲁棒性，在该框架中引入信任对的概念，并相应地对嵌入进行细化，以找到更多的信任对，从而缓解了粗浅学习嵌入所带来的中心问题。文献[74]设计了一种无监督方法用于显微镜图像的无监督分割。首先设计了一个域自适应掩码作为基线，在图像和实例级别进行跨域特征对齐。除了图像级和实例级的域差异之外，上下文信息中还存在语义级的域偏差，并设计一个带有领域鉴别器的语义分割分支，以弥合上下文级别的领域差距。通过整合语义级和实例级的特征自适应，该方法在全景级对跨域特征进行对齐，并在五种数据集中验证了其有效性。文献[75]分层无监督网络对齐方法用于解决物联网时代在不同的网络中相似的物联网设备对齐问题，提出了一种基于循环对抗网

络的无监督网络对齐方法。该方法利用循环对抗网络的对抗特性来实现无监督条件下的实体对齐。

这类方法尤其适用于标注数据稀缺或获取成本高昂的领域,尤其是在大规模数据集的情况下,也能够在不同模态数据间发现潜在、非显性的关联,这些关系可能在有监督的方法中被忽略。同时,由于缺乏明确的指导标签,无监督对齐的结果可能存在较大的不确定性,使得对齐质量在不同应用中有明显差异。某些情况下,无监督对齐方法可能很难达到监督方法在某些任务上的性能水平。

2.1.2 监督方法

监督方法使用有标签的数据来训练模型,以便模型学习如何将来自不同模态的数据进行有效的对齐和整合。这种方法借助人工标注的信息,显式地指导模型捕捉不同模态之间的关联。监督对齐方法常用于需要高准确度对齐的任务,实际应用中可以在无监督对齐技术上进行改进,通过增加模型的监督信息来提高性能。它可以对上述无监督方法进行适当的优化,直接应用于模态对齐任务中。该方法旨在在不降低性能的情况下,尽量减少对监督信息的依赖,也被称为弱监督对齐。文献[76]提出了一种新的有监督的神经自回归模型,以增强学习到的隐藏特征的识别能力,融合多模态数据,通过图卷积网络提取特征,解决了营销意图分析的核心问题。通过加入一定的监督信息,监督方法能够更好地指导模型进行对齐,从而提高对齐的准确性和效果。这种方法可以通过优化无监督方法来实现,以增强模型的性能,也是对无监督学习的进一步发展和延伸。

由于有明确的标签指导对齐过程,监督对齐方法通常能达到较高的性能,相较于无监督方法,监督对齐有明确的优化目标和评价标准,可以更直接高效地训练模型。泛化能力可能不足,原因是在训练数据覆盖的情况下表现优异,但其泛化到未标注或与训练集分布不同的数据上时性能可能下降。当人工标注的数据存在偏差时,将会直接影响模型的准确率。

2.2 隐式对齐

隐式对齐方法是通过模型自身学习来实现不同模态数据之间的对齐,而无需显式指定对齐过程,模态融合的数据集之间的对齐标签数量是巨大的,该方法无疑节省大量人工标注数据标签的成本。在机器翻译的任务中,需要对齐不同语言之间的语义,手工标注工作量大。此时,利用神经网络在模型训练期间对齐不同语言的语义取得了成功的应用。目前最热门的隐式对齐方法是基于注意力机制的对齐方法。它可以有效地识别数据中具有价值的特征区域,通过使用注意力机制,系统可以集中在最相关和有意义的信息上,从而提高任务的性能和效果。这种机制已经成功应用于许多领域。通过引入注意力机制,可以更准确地提取和利用多模态数据中的重要特征,从而提高系统的整体性能。

文献[77]为了将图像与具有语义多样性的多个文本描述对齐,提出一种用于上下文感知的多视图摘要网络,旨在解决视觉形式和文本形式之间的语义鸿沟。文章中用于提取视觉区域和单词的表示的自适应门控自注意力模块,该模块通过控制内部信息流,能够自适应地捕获上下文信息。视觉提取到多视图与提取到的文本信息在注意力机制中计算查询和键之间的关系进行对齐,查询和键可能包含嘈杂无意义的信息。为了更好地对齐有用信息舍弃无用信息,提出自适应门控机制,最终设计多视图匹配模块,将多视图图像特征与相应的文本特征进行匹配。

文献[78]提出 Transformer 编码器推理和对齐网络,使用自注意力模块将视觉信息和文本信息投影到相同的维度空间,用于计算图片和句子的全局相似度矩阵。其不同之处在于采用了最大和池化操作,即计算相似度矩阵每行的最大值并求和,以获取图片和句子的全局相似度,并成功输出细粒度单词区域对齐。

文献[79]提出了一种用于多模态情感识别的门控双向对齐网络。这种方法采用两个独立的卷积神经网络和长短期记忆神经网络编码器,分别从语音和文本中提取特征。它利用基于注意力机制的双向对齐网络来捕捉语音和文本之间的时间相关性。基于文本对齐的语音表示和基于语音对齐的文本表示是从截然不同的模态中获得的,因此基于对齐的表征明显优于使用 BiLSTM 层的最后一个隐藏状态表征。通过门控融合层,能够以可解释的方式自动学习每个表示的贡献,有效地融合多种表示。该网络在情感识别任务中的分类准确性证明了双向对齐网络在提供更具辨别性的情感分类表示方面的有效性。

文献[80]提出基于视觉标记 Transformer 的句子级框架 LipFormer,其中嘴唇运动流、面部标记流的跨模态融合是相互连接的。由自注意力机制产生的两流嵌入到交叉注意模块中,以实现跨视觉和标记变化的对齐。融合后的特征通过级联序列到序列的翻

译解码成语言文本。

虽然基于Transformer的模型在视觉问答领域取得了显著的成功，但它们实现视觉和语言特征对齐的方法简单粗糙。文献[81]提出ST-VQA（shrinkage Transformer-visual question answering）模型，ST-VQA框架使用图像的区域特征作为视觉表示。在不同的Transformer层之间，ST-VQA框架通过特征融合减少了Transformer中的视觉区域数量，并通过对比损失保证了新区域之间的差异，视觉和文本特征被融合并用于决策。该方法比标准的Transformer实现更精准的对齐。

文献[82]提出一种基于有限离散令牌（finite discrete tokens，FDT）的表示，文本信息和图像信息使用各自独立的编码器进行编码，将各自的编码接入到共享的FDT。FDT是一组可学习的标记，用于编码跨模式共享语义概念。图像和文本都表示为模态之间共享的FDT组合，从而使信息粒度统一，进而实现图像与文本语义的对齐。

隐式对齐是在模型内的不同网络层之间巧妙地设计，使得从不同的模态中提取到的特征信息进行映射。例如文献[77]用自注意力机制的查询和键进行匹配对齐；文献[78]在模型内采取顺序堆叠的方式对齐，但这种方式容易匹配成错误的信息；文献[79]采取门控双向的对齐网络用于对齐子模态的对齐；文献[80]不同模态的数据流在交叉注意力模块实现对齐；文献[81]使用收缩的注意力机制可以取得更好的效果。总的来说，隐式对齐方法可以融合没有对齐标签的数据集，可以节省大量的人力标注的成本。通过联合训练或共享表示空间，隐式对齐方法可以实现端到端的学习，直接优化多模态任务的整体性能，也能学习到更加通用的模态对齐，从而提高模型在未见数据上的泛化能力，不足是对齐的质量通常受到模型结构和训练数据的影响，因此其对齐的准确性和稳定性可能难以保证。综上所述，隐式具有自动化、端到端学习和灵活性等优点。其也是未来多模态对齐的主要发展方向。

3 公开数据集

多模态融合的基本是利用不同模态的数据的集成，挖掘不同数据的最大价值。研究者们一直在算法和模型的层面不断地改进和探索，在此基础上不断地取得突破。除了在算法结构和模型方面寻求突破，使用完善数据集也是至关重要的努力方向。这样能够提高基于多模态融合的深度学习模型的性能，并提高输出预测的准确性[83]。多模态数据集也是多个单模态数据集的合集，可以使用多个单模态数据集组合成多模态数据集使用，也可以使用多模态数据集中的部分数据集完成特定任务。下面将详细介绍COCO多模态数据集和CMU-MOSEI多模态数据集。

COCO数据集是在多领域使用的双模态的大规模数据集，经过多次的补充与完善，包含33万张图像，有150万个目标，分成80类不同对象的边界框坐标和完整的分割掩膜和91种无定型的背景区域，每张图像有5条描述语句，且有25万个带关键点标注的行人。利用图像中的150万目标和其类别可以完成目标的检测、部分语义分割等任务。使用图像和每张图像的描述语句可以完成图像标题生成任务[39]。使用带关键点标注的行人可以完成人体姿态估计或者定位行人位置信息等。

CMU-MOSEI数据集包含文本、视觉和声音3种模态信息[31-32]。该数据集收集于You Tube社交平台上的视频。从3 223个视频中分割出23 453视频片段，每个视频片段包含文本语言、手势和面部表情的视觉画面以及语调和韵律的声音。每个句子注释有七分类的情感标注（高度负面、负面、弱负面、中性、弱正面、正面、高度正面）和六分类的情绪标注（高兴、悲伤、生气、恐惧、厌恶、惊讶）。该数据集成为情感识别领域中最经典的数据集，利用该数据集中的文本语言、视觉画面和声音进行多模态融合的情感识别成为非常热门的研究方向。以上数据集能够典型地反映出多模态数据集的特点。本文还列举了一些在多模态融合任务中其他常用的数据集并简述介绍，如表5所示。

4 面临的挑战及未来展望

现有多模态融合技术已经可以显著提升深度学习模型性能，但在融合方面仍有一些问题亟待解决。

（1）多模态数据集目前依旧不够完善，数据集的完善程度将直接影响模型的学习效率与预测准确率的大小，数据量过大，模型参数过多，模型的训练时间过长，引入小样本学习且精度不降低，也是迫切需要解决的问题。下一步将构建更加完善的数据集，对多模态的发展至关重要。

（2）多模态的融合过程中，即在学习互补知识的过程中，很容易引入大量单模或多模的噪声信息，导

表5 多模态融合领域常用经典数据集

Table 5 Classical datasets commonly used in field of multimodal fusion

涉及模态	数据集介绍	参考文献	应用领域
图像+文本	Pascal Sentence：内含1 000张图像，每张图像有5个标注文本语句 CUB-Bird：内含11 788张图片，每张图片有鸟类标记信息	[84][85][86]	匹配与分类
	COCO：内含超33万张图像，含150万目标，每张图像包含5条图像的语句描述	[87]	目标检测 图像标题生成
	Flickr30k：包含30 000张图片和对应的描述句子 Twitter：有24 653图像，有两类别的情感标注	[88][89]	情感分析
	VQAv1.0：在MS-COCO数据集上添加614 000个问题答案对 VQAv2.0：内含265 016图像，平均每张图像5.4个问题和10个基本事实答案 GQA：内含113 018张图像和22 669 678个不同的推理问题	[90][91][92]	视觉问答
	COCO-QA：内含123 287张图像，每张图像有一对问答语句 HoME：内含45 000张3D图像，每张图像都有描述语句和复杂标注 CLEVR：内含100 000张图像，每张图像都有问答语句对	[83][93]	视觉问答
	MSCOCO：内含123 287张图像，每张图像有5个注释文本语句	[94][95]	图片字幕匹配 图像标注
	Flickr8k：内含8 000张图像，每张图像有5个注释文本语句 Flickr30k：内含31 783张图像，每张图像有5个注释文本语句	[77]	
	NUS_WIDE：内含269 648张图像和相关标签，共有5 018个独特标签	[96][97]	
	Conceptual Captions：内含330万个图像-标题对	[98][99]	
	RUC-CAS-WenLan：内含5 500万对图文数据	[100]	
图像+音频	MAHNOB-Mimicry：总计11 h的双人互动视频和音频的记录	[101][102]	视觉问答
	EPIC-Kitchens：内含39 594个动作片段，每个动作都有标注 EPIC-Kitchens-100：内含9万个动作片段累计100 h，每个动作都有标注	[103][104]	物体识别
视频+文本	WebVid-2M：包含200万个视频和文本字幕	[105][106]	多模态检索
	HowTo100M：总计15年时长的23 611个物理世界交互任务，平均时长6.5 min，每段视频平均110视频文本对	[107][108]	
	YT-Temporal-180M：约600万视频段和对应字幕	[109][110]	
视频+音频	IBM AV-ASR Large Vocabulary Studio Dataset	[111]	语音识别
	XM2VTS：内含25位男性和12位女性讲话的声音的视频	[112][113]	
文本+音频	AudioSet：内含2 084 320个10 s音频，总计约5 800 h和527类的人工标注	[114][115]	事件检测
	Aishell-2：约100万对话，总计约1 000 h的音频，包含文本	[116][117]	语义对齐
音频+文本+图像	UR-FUNNY：内含1 866个英文演讲视频及字幕，基于laughter标记的文本数据	[118][119]	幽默检测
	CH-SIMS：内含2 281个视频片段，图像与声音同时出现，都有人工标注的注释	[120]	情绪分类
音频+文本+视频	IEMOCAP：内含488个演员的动作和表情片段，每个演员有9个情感文本标注	[121]	视频情感分类
	MM-IMDB：来自IMDB网站电影的评论，包含正面评论和负面评论各5万条	[122]	影视类型分类
	How2：内含80 000视频片段和单词级别对齐的时间字幕	[123][124]	多模态生成

致出现语义冲突和重复，并最终导致模型过拟合。语义冲突、重复、缺失和噪声等问题仍未有好的解决方法，这些也会影响模型学习的好坏。深入挖掘模态之间的关联关系和互补信息，同时去除冗余信息也是一个难点问题。在多模态融合中注意力机制为融合的首选方法，注意力机制为隐式的融合方法，将数据输入到模型中，就不可进行干预，具有不可解释性。下一步可以构建完善的多模态应用为主的理论体系结构，让模型从多模态数据中发现因果结构并进行定量推断，主动选择可以自我解释的知识，赋予机器智能认知自我推理的能力。

(3)多模态对齐过程中，无监督学习被广泛用于处理无标注数据集，实现数据降维和特征提取；而弱监督学习则更适用于发现不同模态之间的关联关系。监督学习中不同数据集之间需要人工标注，其成本高，有标签的数据集稀缺，导致模型无法进行匹配，迁移学习便可以更好地解决该问题，迁移学习在多模态融合领域的应用可能是未来重要的研究方向之一。

(4)多模态融合结果的量化研究还尚有不足，异

质性的模态如何影响后续的建模，例如不同模态信息融合后对于后续的模型训练，哪种模态更有实质性的贡献，以及在融合之后如何评价不同模态的重要性。因此多模态融合下一步将更深入研究，并提高其在现实应用中的鲁棒性、可解释性和可靠性。

（5）模态噪声拓扑的研究旨在基准测试和改善多模态模型在真实数据不完美情况下的表现。每种模式都有一个独特的噪声拓扑结构，它决定了噪声的分布和它经常遇到的缺陷。例如，图像容易出现模糊和偏移，输入的文本容易在键盘位置后出现打字错误，多模态时间序列数据容易在同步的时间步骤中出现相关缺陷。如何解决该问题也是下一步需要研究的方向。

5 结束语

本文总结了当前深度学习领域中多模态融合技术的最新研究现状，分析阐述了多模态信息融合和多模态信息对齐方面的最新应用。在多模态融合中，根据融合方式的不同，将其分为联合融合方法、协同融合方法、编码器融合方法和分裂融合方法。另一方面，模态对齐作为多模态融合技术的难点，根据数据集是否有监督对齐的标签将对齐方式分为显式对齐和隐式对齐。人工智能领域的研究发展迅速，每一个应用场景都可以使用多模态融合的技术，近期也有大量新型多模态算法的提出，同时也拓展了多模态学习的应用范围，在不同领域的应用中发挥各自的优势和作用。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，有望在今后设计出能够实现与人类智能相匹敌的智能计算机系统。接下来，研究将重点关注多模态组合量化和跨模态迁移学习等问题，旨在推动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] DING M, YANG Z, HONG W, et al. CogView: mastering text-to-image generation via transformers[C]//Advances in Neural Information Processing Systems 34, Dec 6-14, 2021: 19822-19835.

[2] LIU S, FAN H, QIAN S, et al. HiT: hierarchical transformer with momentum contrast for video-text retrieval[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 11915-11925.

[3] MA L, LU Z, LI H. Learning to answer questions from image using convolutional neural network[C]//Proceedings of the 2016 AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2016: 3567-3573.

[4] MARGE M, ESPY-WILSON C, WARD N G, et al. Spoken language interaction with robots: recommendations for future research[J]. Computer Speech & Language, 2022, 71: 101255.

[5] LIANG P P, LYU Y, FAN X, et al. MultiBench: multiscale benchmarks for multimodal representation learning[EB/OL]. [2024-01-06]. https://arxiv.org/abs/2107.07502.

[6] HUANG Y, DU C, XUE Z, et al. What makes multi-modal learning better than single (provably)[C]//Advances in Neural Information Processing Systems 34, Dec 6-14, 2021: 10944-10956.

[7] KARLE P, FENT F, HUCH S, et al. Multi-modal sensor fusion and object tracking for autonomous racing[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(7): 3871-3883.

[8] XIE J, WANG J, WANG Q, et al. A multimodal fusion emotion recognition method based on multitask learning and attention mechanism[J]. Neurocomputing, 2023, 556: 126649.

[9] XU P, ZHU X, CLIFTON D A. Multimodal learning with transformers: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 12113-12132.

[10] LIANG P P, MORENCY L P. Tutorial on multimodal machine learning: principles, challenges, and open questions[C]//Proceedings of the 2023 International Conference on Multimodal Interaction, Paris, Oct 9-13, 2023. New York: ACM, 2023: 101-104.

[11] BALTRUŠAITIS T, AHUJA C, MORENCY L P. Multimodal machine learning: a survey and taxonomy[J]. IEEE Transactions on Pattern analysis and Machine Intelligence, 2018, 41(2): 423-443.

[12] LIU K, FENG G, JIANG X, et al. A feature fusion method for driving fatigue of shield machine drivers based on multiple physiological signals and auto-encoder[J]. Sustainability, 2023, 15(12): 9405.

[13] 王梓衡, 沈继锋, 左欣,等. 基于特征级与决策级融合的农作物叶片病害识别[J]. 江苏大学学报(自然科学版), 2024, 45(3): 286-294.
WANG Z H, SHEN J F, ZUO X, et al. Crop leaf disease recognition based on feature-level and decision-level fusion[J]. Journal of Jiangsu University (Natural Science Edition), 2024, 45(3): 286-294.

[14] HE C, XU P, PEI X, et al. Fatigue at the wheel: a non-visual approach to truck driver fatigue detection by multi-feature fusion[J]. Accident Analysis & Prevention, 2024, 199: 107511.

[15] ZHANG N, WU H, ZHU H, et al. Tomato disease classification and identification method based on multimodal fusion deep learning[J]. Agriculture, 2022, 12(12): 2014.

[16] 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.

[17] WANG M, SHAO W, HUANG S, et al. Hypergraph-regularized multimodal learning by graph diffusion for imaging genetics based Alzheimer's disease diagnosis[J]. Medical Image Analysis, 2023, 89: 102883.

[18] MCFEE B, LANCKRIET G, JEBARA T. Learning multimodal similarity[J]. Journal of Machine Learning Research, 2011, 12(2): 491-523.

[19] 陈辉, 王硕, 许家昌, 等. 基于多尺度特征融合生成对抗网络的水下图像增强[J]. 计算机工程与应用, 2023, 59(21): 231-241.
CHEN H, WANG S, XU J C, et al. Underwater image enhancement based on generate adversarial network with multiscale feature fusion[J]. Computer Engineering and Applications, 2023, 59(21): 231-241.

[20] ZHAO Z, BAI H, ZHU Y, et al. DDFM: denoising diffusion model for multi-modality image fusion[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 8082-8093.

[21] LI Y, QI T, MA Z, et al. Seeking a hierarchical prototype for multimodal gesture recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023. DOI: 10.1109/TNNLS.2023.3295811.

[22] ZHANG H, KOH J Y, BALDRIDGE J, et al. Cross-modal contrastive learning for text-to-image generation[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 833-842.

[23] ZHANG X, HE L, CHEN J, et al. Multiattention mechanism 3D object detection algorithm based on RGB and LiDAR fusion for intelligent driving[J]. Sensors, 2023, 23(21): 8732.

[24] CHEN J, HU Y, LAI Q, et al. IIFDD: intra and inter-modal fusion for depression detection with multi-modal information from Internet of medical things[J]. Information Fusion, 2024, 102: 102017.

[25] SINGH A, HU R, GOSWAMI V, et al. FLAVA: a foundational language and vision alignment model[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 15638-15650.

[26] LI J, LI D, SAVARESE S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models[C]//Proceedings of the 2023 International Conference on Machine Learning, Honolulu, Jul 23-29, 2023: 19730-19742.

[27] LIANG S, ZHAO M, SCHÜTZE H. Modular and parameter-efficient multimodal fusion with prompting[EB/OL]. [2024-01-06]. https://arxiv.org/abs/2203.08055.

[28] KUMARI R, EKBAL A. AMFB: attention based multimodal factorized bilinear pooling for multimodal fake news detection [J]. Expert Systems with Applications, 2021, 184: 115412.

[29] XU Q, MEI Y, LIU J, et al. Multimodal cross-layer bilinear pooling for RGBT tracking[J]. IEEE Transactions on Multimedia, 2021, 24: 567-580.

[30] GOEL T, SHARMA R, TANVEER M, et al. Multimodal neuroimaging based Alzheimer's disease diagnosis using evolutionary RVFL classifier[J]. IEEE Journal of Biomedical and Health Informatics, 2023, 6: 1-9.

[31] HAN G, WANG M, ZHU H, et al. UIEGAN: adversarial learning-based photo-realistic image enhancement for intelligent underwater environment perception[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5611514.

[32] ZHAO C, YANG P, ZHOU F, et al. MHW-GAN: multi-discriminator hierarchical wavelet generative adversarial network for multimodal image fusion[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023. DOI: 10.1109/TNNLS.2023.3271059.

[33] ELAKKIYA R, VIJAYAKUMAR P, KUMAR N. An optimized generative adversarial network based continuous sign language classification[J]. Expert Systems with Applications, 2021, 182: 115276.

[34] YANG B, XIANG X, KONG W, et al. DMF-GAN: deep multimodal fusion generative adversarial networks for text-to-image synthesis[J]. IEEE Transactions on Multimedia, 2024, 26: 6956-6967.

[35] FAN H, ZHANG X, XU Y, et al. Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals[J]. Information Fusion, 2024, 104: 102161.

[36] LIU Y, BING W, REN S, et al. BC-FND: an approach based on hierarchical bilinear fusion and multimodal consistency for fake news detection[J]. IEEE Access, 2024, 12: 62738-62749.

[37] KANG B, LIANG D, MEI J, et al. Robust RGB-T tracking via graph attention-based bilinear pooling[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34 (12): 9900-9911.

[38] DAS R, SINGH T D. Image-text multimodal sentiment analysis framework of assamese news articles using late fusion[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2023, 22(6): 1-30.

[39] 黄忠, 胡敏, 刘娟. 基于多特征决策级融合的表情识别方法[J]. 计算机工程, 2015, 41(10): 171-176.

HUANG Z, HU M, LIU J. Facial expression recognition method based on multi-feature decision-level fusion[J]. Computer Engineering, 2015, 41(10): 171-176.

[40] 宁大海, 郑晟. 可见光和红外图像决策级融合目标检测算法[J]. 红外技术, 2023, 45(3): 282-291.
NING D H, ZHENG S. An object detection algorithm based on decision-level fusion of visible and infrared images[J]. Infrared Technology, 2023, 45(3): 282-291.

[41] HAN W, CHEN H, GELBUKH A, et al. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis[C]//Proceedings of the 2021 International Conference on Multimodal Interaction, Montréal, Oct 18-22, 2021. New York: ACM, 2021: 6-15.

[42] YANG B, WU L, ZHU J, et al. Multimodal sentiment analysis with two-phase multi-task learning[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 2015-2024.

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

[44] RASIWASIA N, COSTA PEREIRA J, COVIELLO E, et al. A new approach to cross-modal multimedia retrieval[C]//Proceedings of the 18th ACM International Conference on Multimedia. New York: ACM, 2010: 251-260.

[45] FROME A, CORRADO G S, SHLENS J. DeViSE: a deep visual-semantic embedding model[C]//Advances in Neural Information Processing Systems 26, Lake Tahoe, Dec 5-8, 2013: 2121-2129.

[46] MEKHALDI D. Multimodal document alignment: towards a fully-indexed multimedia archive[C]//Proceedings of the 2007 Multimedia Information Retrieval Workshop, Amsterdam, Jul 23-27, 2007.

[47] WEHRMANN J, MATTJIE A, BARROS R C. Order embeddings and character-level convolutions for multimodal alignment[J]. Pattern Recognition Letters, 2018, 102: 15-22.

[48] SONG G, WANG S, TIAN Q. Fusing feature and similarity for multimodal search[C]//Proceedings of the 2015 IEEE China Summit and International Conference on Signal and Information Processing. Piscataway: IEEE, 2015: 787-791.

[49] HU D, NIE F, LI X. Deep multimodal clustering for unsupervised audiovisual learning[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 9248-9257.

[50] LIANG P P, ZADEH A, MORENCY L P. Foundations & trends in multimodal machine learning: principles, challenges, and open questions[J]. ACM Computing Surveys, 2024, 56 (10): 264.

[51] VENDROV I, KIROS R, FIDLER S, et al. Order-embeddings of images and language[EB/OL]. [2024-01-06]. https://arxiv.org/abs/1511.06361.

[52] YOUNG P, LAI A, HODOSH M, et al. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions[J]. Transactions of the Association for Computational Linguistics, 2014, 2: 67-78.

[53] MAJUMDER N, HAZARIKA D, GELBUKH A, et al. Multimodal sentiment analysis using hierarchical fusion with context modeling[J]. Knowledge-Based Systems, 2018, 161: 124-133.

[54] QU Z, WANG C Y, WANG S Y, et al. A method of hierarchical feature fusion and connected attention architecture for pavement crack detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(9): 16038-16047.

[55] LI R, WU X, LI A, et al. HFBSurv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction[J]. Bioinformatics, 2022, 38(9): 2587-2594.

[56] WANG J, YANG S, ZHAO H. Crisis event summary generative model based on hierarchical multimodal fusion[J]. Pattern Recognition, 2023, 144: 109890.

[57] ZHOU W, DONG S, LEI J, et al. MTANet: multitask-aware network with hierarchical multimodal fusion for RGB-T urban scene understanding[J]. IEEE Transactions on Intelligent Vehicles, 2022, 8(1): 48-58.

[58] 任泽裕, 王振超, 柯尊旺, 等. 多模态数据融合综述[J]. 计算机工程与应用, 2021, 57(18): 49-64.
REN Z Y, WANG Z C, KE Z W, et al. Survey of multimodal data fusion[J]. Computer Engineering and Applications, 2021, 57(18): 49-64.

[59] NIE L, WANG W, HONG R, et al. Multimodal dialog system: generating responses via adaptive decoders[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 1098-1106.

[60] RICHARD A, LEA C, MA S, et al. Audio-and gaze-driven facial animation of codec avatars[C]//Proceedings of the 2021 IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2021: 41-50.

[61] TSAI Y H H, LIANG P P, ZADEH A, et al. Learning factorized multimodal representations[EB/OL]. [2024-01-06]. https://arxiv.org/abs/1806.06176.

[62] SHI Y, PAIGE B, TORR P. Variational mixture-of-experts autoencoders for multi-modal deep generative models[C]//Advances in Neural Information Processing Systems 32, Vancouver, Dec 8-14, 2019: 15692-15703.

[63] CHEN B, ROUDITCHENKO A, DUARTE K, et al. Multimodal clustering networks for self-supervised learning from

unlabeled videos[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 8012-8021.

[64] WANG L, QIAO Y, TANG X. Action recognition with trajectory-pooled deep-convolutional descriptors[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2015: 4305-4314.

[65] KARPATHY A, FEI-FEI L. Deep visual-semantic alignments for generating image descriptions[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2015: 3128-3137.

[66] SRIVASTAVA N, SALAKHUTDINOV R R. Multimodal learning with deep Boltzmann machines[C]//Advances in Neural Information Processing Systems 25, Lake Tahoe, Dec 3-6, 2012: 2231-2239.

[67] TAPASWI M, BÄUML M, STIEFELHAGEN R. Aligning plot synopses to videos for story-based retrieval[J]. International Journal of Multimedia Information Retrieval, 2015, 4: 3-16.

[68] TAPASWI M, BAUML M, STIEFELHAGEN R. Book2movie: aligning video scenes with book chapters[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2015: 1827-1835.

[69] HUDDAR M G, SANNAKKI S S, RAJPUROHIT V S. Attention-based word-level contextual feature extraction and cross-modality fusion for sentiment analysis and emotion classification[J]. International Journal of Intelligent Engineering Informatics, 2020, 8(1): 1-18.

[70] LI H, DING W, WU Z, et al. Learning fine-grained cross modality excitement for speech emotion recognition[EB/OL]. [2024-01-06]. https://arxiv.org/abs/2010.12733.

[71] CHEN B, CAO Q, HOU M, et al. Multimodal emotion recognition with temporal and semantic consistency[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3592-3603.

[72] LI K, CHEN C, CAO W, et al. DeAF: a multimodal deep learning framework for disease prediction[J]. Computers in Biology and Medicine, 2023, 156: 106715.

[73] SUN Q, LIN X, ZHANG Y, et al. Towards higher-order topological consistency for unsupervised network alignment[C]//Proceedings of the 2023 IEEE 39th International Conference on Data Engineering. Piscataway: IEEE, 2023: 177-190.

[74] LIU D, ZHANG D, SONG Y, et al. PDAM: a panoptic-level feature alignment framework for unsupervised domain adaptive instance segmentation in microscopy images[J]. IEEE Transactions on Medical Imaging, 2020, 40(1): 154-165.

[75] ZHU D, SUN Y, DU H, et al. HUNA: a method of hierarchical unsupervised network alignment for IoT[J]. IEEE Internet of Things Journal, 2020, 8(5): 3201-3210.

[76] ZHANG L, SHEN J, ZHANG J, et al. Multimodal marketing intent analysis for effective targeted advertising[J]. IEEE Transactions on Multimedia, 2021, 24: 1830-1843.

[77] QU L, LIU M, CAO D, et al. Context-aware multi-view summarization network for image-text matching[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 1047-1055.

[78] MESSINA N, AMATO G, ESULI A, et al. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2021, 17(4): 1-23.

[79] LIU P, LI K, MENG H, et al. Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition[J]. Neurocomputing, 2022, 496: 46-55.

[80] XUE F, LI Y, LIU D, et al. LipFormer: learning to lipread unseen speakers based on visual-landmark transformers[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(9): 4507-4517.

[81] XIA H, LAN R, LI H, et al. ST-VQA: shrinkage transformer with accurate alignment for visual question answering[J]. Applied Intelligence, 2023, 53(18): 20967-20978.

[82] CHEN Y, YUAN J, TIAN Y, et al. Revisiting multimodal representation in contrastive learning: from patch and token embeddings to finite discrete tokens[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 15095-15104.

[83] 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.
LIANG B, LIU Q, XU J, et al. Aspect-based sentiment analysis based on multi-attention CNN[J]. Journal of Computer Research and Development, 2017, 54(8): 1724-1735.

[84] CHENG Q, TAN Z, WEN K, et al. Semantic pre-alignment and ranking learning with unified framework for cross-modal retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022. DOI: 10.1109/TCSVT.2022.3182549.

[85] LIAO L, YANG M, ZHANG B. Deep supervised dual cycle adversarial network for cross-modal retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(2): 920-934.

[86] LIAO W, HU K, YANG M Y, et al. Text to image generation with semantic-spatial aware GAN[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern recognition. Piscataway: IEEE, 2022: 18187-18196.

[87] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Proceedings of the 13th European Conference on Computer Vision, Zurich, Sep 6-12, 2014. Cham: Springer, 2014: 740-755.

[88] ZHAO D, CHANG Z, GUO S. A multimodal fusion approach for image captioning[J]. Neurocomputing, 2019, 329: 476-485.

[89] BIBI M, ABBASI W A, AZIZ W, et al. A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for Twitter sentiment analysis[J]. Pattern Recognition Letters, 2022, 158: 80-86.

[90] HUDSON D A, MANNING C D. GQA: a new dataset for real-world visual reasoning and compositional question answering [C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 6700-6709.

[91] GAO Y, CAO Y, KOU T, et al. VDPVE: VQA dataset for perceptual video enhancement[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 1474-1483.

[92] YANG G, ZHANG Z, LIU X. Visual question answering model based on fusing global-local feature[C]//Proceedings of the 3rd International Conference on Computer Vision and Pattern Analysis, Hangzhou, Mar 31-Apr 2, 2023: 6-11.

[93] YANG Z, XIANG J, YOU J, et al. Event-oriented visual question answering: the E-VQA dataset and benchmark[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(10): 10210-10223.

[94] ZHANG Z, ZHANG Y, ZHANG Y, et al. Vital information is only worth one thumbnail: towards efficient human pose estimation[J]. Pattern Recognition, 2024, 147: 110111.

[95] CHENA Y, LIUA J, YANG Z, et al. Active mining sample pair semantics for image-text matching[EB/OL]. [2024-01-06]. https://arxiv.org/abs/2311.05425.

[96] CHUA T S, TANG J, HONG R, et al. NUS-WIDE: a real-world web image database from National University of Singapore [C]//Proceedings of the 2009 ACM International Conference on Image and Video Retrieval. New York: ACM, 2009: 1-9.

[97] GUPTA A, NARAYAN S, KHAN S, et al. Generative multi-label zero-shot learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 14611-14624.

[98] HU X, GAN Z, WANG J, et al. Scaling up vision-language pre-training for image captioning[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 17980-17989.

[99] CHEN J, GUO H, YI K, et al. VisualGPT: data-efficient adaptation of pretrained language models for image captioning[C]// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 18030-18040.

[100] HUO Y, ZHANG M, LIU G, et al. WenLan: bridging vision and language by large-scale multi-modal pre-training[EB/OL]. [2024-01-06]. https://arxiv.org/abs/2103.06561.

[101] TUYEN N T V, GEORGESCU A L, DI GIULIO I, et al. A multimodal dataset for robot learning to imitate social human-human interaction[C]//Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. New York: ACM, 2023: 238-242.

[102] VILCHIS C, GONZALEZ-MENDOZA M, CHANG L, et al. A study of the frameworks for digital humans: analyzing facial tracking evolution and new research directions with AI[C]//Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Feb 6-8, 2022: 154-162.

[103] DAMEN D, DOUGHTY H, FARINELLA G M, et al. The EPIC-Kitchens dataset: collection, challenges and baselines [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(11): 4125-4141.

[104] HUANG Z, QING Z, WANG X, et al. Towards training stronger video vision transformers for EPIC-Kitchens-100 action recognition[EB/OL]. [2024-01-06]. https://arxiv.org/abs/2106.05058.

[105] WANG J, GE Y, CAI G, et al. Object-aware video-language pre-training for retrieval[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 3313-3322.

[106] SHI Y, LIU H, XU H, et al. Learning semantics-grounded vocabulary representation for video-text retrieval[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 4460-4470.

[107] HUANG P Y, PATRICK M, HU J, et al. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models[EB/OL]. [2024-01-06]. https://arxiv.org/abs/2103.08849.

[108] HAN T, XIE W, ZISSERMAN A. Temporal alignment networks for long-term video[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 2906-2916.

[109] CHEN S, LI H, WANG Q, et al. VAST: a vision-audio-subtitle-text omni-modality foundation model and dataset[C]//Advances in Neural Information Processing Systems 36, New Orleans, Dec 10-16, 2023.

[110] YANG Z, FANG Y, ZHU C, et al. i-Code: an integrative and composable multimodal learning framework[C]//Proceedings of the 2023 AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2023: 10880-10890.

[111] GAONKAR A, CHUKKAPALLI Y, RAMAN P J, et al. A comprehensive survey on multimodal data representation

and information fusion algorithms[C]//Proceedings of the 2021 International Conference on Intelligent Technologies. Piscataway: IEEE, 2021: 1-8.

[112] TORRIE S, SUMSION A, SUN Z, et al. Automated dataset collection pipeline for lip motion authentication[J]. Electronic Imaging, 2023, 35(5).

[113] RADMAN A, SALLAM A, SUANDI S A. Deep residual network for face sketch synthesis[J]. Expert Systems with Applications, 2022, 190: 115980.

[114] FONSECA E, FAVORY X, PONS J, et al. Fsd50k: an open dataset of human-labeled sound events[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 30: 829-852.

[115] CHONG D, WANG H, ZHOU P, et al. Masked spectrogram prediction for self-supervised audio pre-training[C]//Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2023: 1-5.

[116] WANG M, CHEN J, ZHANG X L, et al. End-to-end multimodal speech recognition on an air and bone conducted speech corpus[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 31: 513-524.

[117] ZHOU X, WANG J, CUI Z, et al. MMSpeech: multi-modal multi-task encoder-decoder pre-training for speech recognition[EB/OL]. [2024-01-06]. https://arxiv.org/abs/2212.00500.

[118] ZENG Y, LI Z, CHEN Z, et al. A feature-based restoration dynamic interaction network for multimodal sentiment analysis [J]. Engineering Applications of Artificial Intelligence, 2024, 127: 107335.

[119] KIM K, PARK S. AOBERT: all-modalities-in-one BERT for multimodal sentiment analysis[J]. Information Fusion, 2023, 92: 37-45.

[120] ZHANG L, LIU C, JIA N. Uni2mul: a conformer-based multimodal emotion classification model by considering unimodal expression differences with multi-task learning [J]. Applied Sciences, 2023, 13(17): 9910.

[121] REN M, HUANG X, LIU J, et al. MALN: multimodal adversarial learning network for conversational emotion recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(11): 6965-6980.

[122] SEO S B, NAM H, DELGOSHA P. MM-GATBT: enriching multimodal representation using graph attention network[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop. Stroudsburg: ACL, 2022: 106-112.

[123] LIU N, WEI K, SUN X, et al. Assist non-native viewers: multimodal cross-lingual summarization for how2 videos[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2022: 6959-6969.

[124] LIU N, SUN X, YU H, et al. Abstractive summarization for video: a revisit in multistage fusion network with forget gate[J]. IEEE Transactions on Multimedia, 2023, 25: 3296-3310.

张虎成(1998—),男,硕士研究生,主要研究方向为数据处理、多模态融合等。

ZHANG Hucheng, born in 1998, M.S. candidate. His research interests include data processing, multimodal fusion, etc.

李雷孝(1978—),男,博士,教授,CCF会员,主要研究方向为数据分析与数据挖掘、网络空间安全、区块链技术等。

LI Leixiao, born in 1978, Ph.D., professor, CCF member. His research interests include data analysis and data mining, cyberspace security, block chain technology, etc.

刘东江(1988—),男,博士,主要研究方向为数据挖掘、图计算、机器学习、区块链。

LIU Dongjiang, born in 1988, Ph.D. His research interests include data mining, graph computing, machine learning and blockchain.