拉马钱德兰姆和
格雷厄姆·W。泰勒股份有限公司

深度多模态学习

关于最新进展和趋势的调查

©Istockphoto.com/zapp2pho

T深度学习的成功一直是解决日益复杂的机器学习问题的催化剂，这些问题通常涉及多种数据模式。我们回顾了深度多模态学习的最新进展，并强调了最新的技术现状，以及这一活跃的研究领域的差距和挑战。我们首先对深度多模态学习架构进行分类，然后讨论在深度学习架构中融合学习到的多模态表示的方法。我们强调了两个研究的领域-正则化策略和方法，学习或优化多模态融合结构-作为令人兴奋的领域未来的工作。

产品介绍

近年来，神经网络取得了令人印象深刻的复苏，长期以来，他们利用算法、数据和计算[1]方面的进步，成功地减轻了人们对训练深度模型能力的担忧。这个活跃的研究领域现在引起了学术界的兴趣，但也引起了工业界的兴趣，它为许多实际问题取得了最先进的性能，特别是在涉及高维非结构化数据的领域，如计算机视觉、语音和自然语言处理。

随着深度学习在视觉领域取得了不可否认的成功，深度学习研究的自然进展指向了涉及更大、更复杂的多模态数据的问题。这种多模态数据集由来自观察到一个共同现象的不同传感器的数据组成，其目标是以一种互补的方式使用这些数据来学习一个复杂的任务。深度学习的主要优点之一是，可以自动学习每种模态的层次表示，而不是手动设计或手工制作特定于模态的特征，然后将其输入机器学习算法。

本文的目标是对深入多模态学习的最先进水平进行全面的调查，并通过突出这个活跃领域的进展、差距和挑战，提出未来的研究方向。我们认为，鉴于深度学习技术数量的增加，这篇综述是及时的

数字对象标识符10.1109/MSP.2017.2738401发布日期：2017年11月13日

应用于在主要会议和期刊上发表的多模式数据，如图1所示。

本文的关键集中在深度多模态学习研究的两个重要领域：1)使用正则化技术改进跨模态学习的方法（见“多模态正则化”部分)和2)试图通过搜索、优化或某些学习过程找到最佳深度多模态体系结构的方法(参见“融合结构学习和优化”部分）。

背景背景

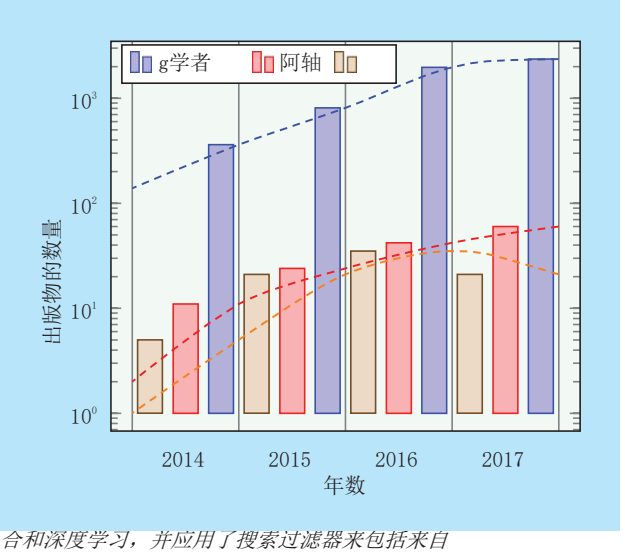
为了进行我们的审查，我们采用了Lahat等人提供的定义。[2]，其中我们考虑使用多个传感器观察到的现象或系统，每个传感器输出可以被称为与单个数据集相关的模态。使用多模态数据的潜在动机是，可以从给定学习任务考虑的每种模式中提取互补信息，与只使用单一表示相比，可以使用产生大大改进的性能。多模态数据的使用受益于许多实际任务。例如，在医学图像分析中，使用多种成像方式，如计算机断层扫描(CT)、磁共振成像(MRI)和超声成像，提供了医学专家在诊断和治疗中常规使用的补充信息。涉及人机交互的应用程序广泛使用沉浸式游戏和自动驾驶等应用程序的深度和视觉线索。在生物识别应用程序中也可以看到类似的性能改进。在遥感应用中，来自不同传感器（强度图像、合成孔径雷达和光探测和测距（激光雷达））的数据经常被融合。

研究界[3]，[4]长期研究多模态数据融合技术，涵盖不同的应用领域。传统上，结合多个传感器的信号是从数据融合的角度来研究的。这被称为早期融合或数据级融合，并专注于如何最好地组合来自多个来源的数据，或通过去除模式之间的相关性，或在一个低维的公共子空间中表示融合的数据。实现其中一个或两个目标的技术包括主成分分析(PCA)、独立成分分析和规范相关分析。然后将融合的数据呈现给一个机器学习算法。*当集成分类器在21世纪初[5]流行时，研究人员开始应用多模态融合技术，被称为晚期融合或决策级融合。*一般来说，这些晚期融合策略比早期融合更容易实现得多，特别是当不同的模式在数据维数和采样率方面存在显著差异，并往往导致性能的提高时。

正如“中间融合”一节所示，流行的深度神经网络(DNN)结构允许第三种多模态融合，即学习表征的中间融合，为许多实际问题的多模态融合提供了一种真正灵活的方法。作为深度学习架构学习底层数据隐藏层中底层数据的层次表示，不同模式之间的学习表示可以在不同的抽象层次上融合。

与传统的机器学习方法相比，基于深度学习的多模态学习有几个优势，如表1所示。对于许多实际问题，深度学习模型通常为涉及多模态数据的问题提供很多改进的性能。然而，这需要我们接下来要讨论的几个体系结构设计选择。

第一个设计选择与何时融合不同的模式有关。从传统的数据融合的角度来看，从业者可以在数据级别融合各种输入模式，并继续训练单一的机器学习模型，但是，正如我们在“早期融合”一节中讨论的，这个选项可能相当具有挑战性。或者，也可以考虑一个晚期融合选项，我们在“晚期融合”一节中回顾了这一类别中的一些作品。深度学习的一个重要特征是它能够从原始数据中学习层次表示。这一特性可以用于多模态学习，以对学习到的表示如何融合进行细粒度控制。因此，多模态深度学习的一个常见做法是构建一个共享的表示或融合层，可以合并模态的传入表示，从而迫使网络学习其输入的联合表示。最简单的融合层是一层隐藏单元，每一个单元都接收来自所有模式的输入。在不同的抽象层次上学习跨模态共享表示的灵活性可以用来实现更好的多模态融合结果；然而，



合和深度学习，并应用了搜索过滤器来包括来自
工程、计算机科学和数学领域，不包括与社会科学、神经生物学和商业相关的结果。注意，由于返回结果的规模差异，我们使用了半日志量表。

表1。深度多模态学习与传统方法的比较。

深度多模态学习	传统的多模态学习
学习数据的模式表示（特征）和共享（融合）表示。	特性是手动设计的，需要关于潜在问题和数据的先验知识。
很少或不需要对输入数据的预处理（端到端培训）。	一些技术，如早期融合，可能对数据预处理很敏感。实现架构中
的隐式降维功能。	特征选择和降维通常被明确地执行。支持早期、晚期或中间阶段
的融合。	通常进行早期或晚期的融合。
易于扩展的数据大小和模式的数量的所有融合方法。	早期融合（数据级融合）可能具有挑战性，而且不可扩展：可能需要定义后期融合规则。
融合架构可以在学习过程中学习。	刚性融合架构通常是手工制作的。
更深、更复杂的网络通常需要大量的训练数据（如果从头开始训练）。	可能不需要那么多的培训数据。
许多超参数调谐对最先进的性能至关重要。	可能没有深度学习架构有那么多的超参数。
计算密集型的用户，需要强大的图形处理单元(gpu)才能获得合理的训练时间。	gpu可能会提供加速功能，但并不是至关重要的。

问题仍然存在：在何种深度的表示形式下，融合将是最佳的？

深度多模态学习的第二种架构设计选择涉及到哪种模式需要融合。多模态融合的潜在假设是，不同的模式为解决手头的任务提供了互补的信息。然而，可能的情况是，包含所有可用的模式最终不利于机器学习算法的性能——因此，可能需要某种形式的特征选择。在“融合结构学习和优化”一节中，我们将讨论许多技术，在训练过程中，可以自动学习最佳顺序和深度的融合。第三种设计选择涉及到处理缺失的数据或模式。深度多模态学习模型应该足够健壮，以补偿推理过程中缺失的数据或模式。生成模型通常用于这种情况下。大多数深度多模态学习方法也涉及到从原始数据中进行的表示学习。通常，深度多模模式体系结构利用了针对特定类型的数据进行优化的多个标准模块或“构建块”。选择哪些深度学习模块最适合提取给定模态的相关信息，这也是一个重要的架构设计选择。例如，当考虑到基于二维（二维）像素的数据时，卷积体系结构通常是首选的。三维（三维）卷积网络可以用于体积数据，如CT、MRI，甚至视频。当使用时间数据时，循环神经网络(RNNs)的变体，如长期短期记忆(LSTM)或可以合并门控循环单元。

模态级深度学习架构的选择主要取决于输入的维数或是否需要学习时间趋势。除了这些常见的体系结构选择之外，它还应由读者来决定，给定特定于应用程序的需求，这可能涉及数据集的属性，甚至是用于培训或部署的硬件。

应用程序

本节旨在概述深度多模态学习引起极大兴趣的各种应用领域。虽然多模态学习和融合是一个被广泛研究的课题，但深度多模态学习是在Ngiam等人的研究下才开始受到关注的。[6]和斯里瓦斯塔瓦和萨拉库迪诺夫[7]。这些早期关于深度多模态融合的工作只涉及两种模式：图像和文本。Ngiam等人。[6]研究了几种多模态融合的方法，包括简单的输入连接和共享表示学习，以及交叉模态学习（在训练过程中存在来自所有模态的数据，但在测试过程中只有一个单一的模态可用）。大约在同一时间，斯里瓦斯塔瓦和萨拉克特迪诺夫[7]也展示了在深度学习框架中融合涉及图像和文本的不同模式的更高层次表示的效用。一个值得注意的发现是，通过简单的传入连接的连接来构建一个多模态融合层会导致相对较差的结果——这表明隐藏单元与来自单个模式的变量有很强的连接，但很少有单元跨模式连接。他们还发现，捕获交叉模态相关性需要至少一个非线性阶段才能成功，因为单个模态的更高水平表示将相对“无模态”，因此更容易融合。这些早期的探索成为了深度多模态学习中许多继续工作的基础，这些工作研究了各种正则化策略（见“多模态正则化”部分），以加强对学习模态间关系的约束。

人类活动识别

大量利用多模态数据的一个重要研究领域是人类活动识别。在这个庞大的研究保护伞下，有许多研究的子领域与人类理解的某些方面有关。鉴于那个人类

表2。多模态学习数据集和公共多模态机器学习的挑战。

数据集	操作模式	出现的问题	参考资料	年公
司-MHAD	深度和惯性传感器数据	人类的行动识别	陈等人。[93]	2015
观察人的人	RGB-D，音频，骨骼姿势	人类活动识别	天秤座等人。[94]	2014
伯克利大学MHAD	多视点RGB-D和骨骼姿势数据	人类活动识别	Ofli等人。[95]	2013
MHRI数据集	胸部，顶部RGB-D，面部，视频，和音频	人机交互作用	巴勃罗等人。[96]	2016
氢电机	9个智能手机传感器和交互数据	在智能手机中的持续身份验证	Sitova等人。[12]	2016
复拉	音频、视觉和生理功能	情绪识别	林格瓦尔等人。[97]	2013
移动设备的健康状况	加速度计、心电图、磁强计和陀螺仪	健康状况监测	巴诺斯等人。[98]	2015
不感兴趣的多模态的	图像和文字（40米）	多模态的单词嵌入	毛等人。[99]	2016
毫米-IMDb	视频、图像和文本元数据	电影类型的预测	阿雷瓦洛等人。[100]	2017
fcvid	视频和音频	行动识别	姜江等人。[101]	2017
基蒂岛	立体声灰度和彩色视频，3d激光雷达，惯性和GPS导航数据	自动驾驶	盖格尔等人。[91]	2017
驱动器tFaceDB	RGB-D和面部标志物	人脸识别	Min等人。[102]	2014
牛津机器人驾驶车	六台摄像机、激光雷达、GPS和惯性导航数据	自动驾驶	马德恩等人。[92]	2016
多模态的孩子	T2-、氟、钆后T1-MRI、灌注、弥散MRI和MRSI	脑肿瘤的分割	Menze等人。[103]	2015

RGB-D：RGB-深度

在社会环境中表现出高度复杂的行为，机器学习算法需要多模态数据来分类或“理解”其人类行为是很自然的。毫不奇怪，我们发现近年来报道的许多深度多模态融合的工作都集中在多媒体数据上，这些数据通常涉及音频、视频、深度和骨骼运动信息。多模态深度学习方法已被应用于涉及人类活动的各种问题，如动作识别（一个活动可以由两个或两个以上更短的动作序列组成）、手势识别[8]、注视方向估计[9]、人脸识别[10]和情绪识别[11]。移动智能手机拥有不少于10个传感器，这已经引发了涉及多模态数据的新应用，如连续生物特征认证[12]和活动识别[13]。相关的研究子领域包括人体姿态估计[14]和语义场景理解[15]。

我们预计，在可预见的未来，深度学习研究社区将继续关注这些问题。这不仅可以从在线发表的多模模式阅读论文的数量，而且可以通过越来越多的数据集和公共挑战来证明这一点（见表2）。

医疗用应用程序

医学应用中的深度学习已成为引起人们广泛兴趣的重要应用领域。例如，医学成像，由许多以不同医学成像方式的形式提供的多模态数据，如MRI、CT、正电子发射断层扫描(PET)、功能MRI(fMRI)、x射线和超声。虽然新的医学成像技术已经有了显著的改进，但对这些诊断方式的解释仍然需要高度训练有素的人类专家。传统的计算机视觉方法需要手动设计的形态学特征表示。然而，将人类专家的隐性知识转化为一种计算形式并不是一件小事。在医学成像领域，手动设计合适的图像特征是极具挑战性的，因为它不仅涉及解释微妙的视觉标记和异常，通常需要多年的医学培训，而且还需要融合来自多种成像模式的互补和可能相互冲突的信息。因此，通过例子学习这些多模态特征的能力，如深度学习应用于计算机视觉的成功情况，促使研究人员研究它们在医学领域的适用性。因此，近年来越来越多的医学图像分析研究，[16]，无论是单模态还是多模态，都涉及到基于深度学习的方法，也就不足为奇了。

多模态深度学习已被用于组织和器官分割[17]、多模态医学图像检索[18]、多模态医学图像配准[19]和计算机辅助诊断[20]等问题。马莫什纳等人最近的一篇综述文章。[21]演示了

dnn的日益流行，包括在涉及基因组、蛋白质组学和药物数据的生物医学应用中实现多模态融合的模型。

在医学应用中，基于深度学习的方法应用的两个主要挑战是1)获得足够标记数据的困难和2)类不平衡的问题（消极例子的数量远远超过积极例子的数量）。为了克服第一个挑战，早期的方法采用了基于补丁的训练[22]。最近，利用迁移学习的技术取得了惊人的成功。这涉及到重用由非常深的、大型数据集上的网络学习到的部分数据不可知表示，例如，ImageNet，然后只使用小得多的医疗数据集对网络的上层进行微调或再训练。另一种常见的技术是执行训练数据增强，例如，应用不同的仿射变换或随机干扰图像的亮度和对比度，以增加可用的训练数据量。为了解决数据不平衡问题，通常会对损失函数应用某种形式的加权，这样在大多数类上犯的错误比网络在少数类上犯的错误受到的惩罚要小。这些挑战虽然在医学领域的问题中很常见，但也可能发生在其他领域，因此，解决方案也同样适用。然而，尽管深度学习在医学应用中取得了成功，但医学界仍然相当担心在现实世界中部署它们，因为深度学习通常被视为不透明的。考虑到设计可解释dnn[24]，[25]的不断努力，这种观点可能会逐渐改变。

自主的系统

随着深度学习的成功，人们对自主导航（也被称为自动驾驶）应用程序的兴趣激增，这些应用程序通常涉及从安装在车辆上的传感器中获得的多模态数据。例如，自动驾驶汽车可以包括许多外部和内部传感器，包括雷达、立体可见光相机、激光雷达、红外(IR)相机、全球定位系统(GPS)和音频。为了执行自主导航，从传感器收集的异构数据用于学习一些相关但复杂的任务，如本地定位和映射、场景理解、运动规划和驾驶员状态识别。

自主导航面临的最大挑战之一是运行环境的动态特性——该系统必须适应并对天气变化、照明变化、行人和其他交通、道路状况、交通标志以及驾驶员的状态做出反应。尽管如此，低盈利和强化学习技术[26]在优步、英伟达、百度和特斯拉等行业参与者中积极参与商业自动驾驶汽车的开发，推动了这一应用领域的发展。

实时了解自动驾驶的场景是一个重要的任务。它需要学习系统来识别场景中的物体，如车道、交通标志、行人和其他交通。因此，对于多模态视频馈送的每一帧，语义分割必须首先是已执行的操作。在场景中标识的每个语义概念都必须进行本地化。对于这类任务，通常使用对每个帧执行像素级标记的深度完全卷积体系结构的[27]。对于多模态输入，一种常见的策略是在被输入到完全卷积神经网络(CNN)（在某种意义上是早期融合）之前，跨通道维度连接同步帧，或者，训练单独的模态网络，然后在多模态网络的更深层次阶段融合。我们在“融合结构”一节中进一步讨论了这种融合策略。“语义分割可以通过使用全cnn的三维变体来扩展到视频。自驾驶汽车技术中使用的基本技术可以扩展到其他机器人应用，如移动机器人或无人机导航、抓取配置学习[28]和机器人操作[29]。

总结

我们强调了深度多模态学习方法获得立足点并继续经历快速发展的三个主要领域。除了已经强调的关于这些关键领域的工作外，我们还在表3中列出了涉及深度多模态学习的其他代表性工作。其他几个涉及文本、图像和视频的应用程序领域，例如，视觉问题回答(VQA)和图像和视频注释，将在接下来的章节中突出显示，我们将讨论特定的深度学习模型。

模型

将多模态深度学习应用于一个新的问题中，涉及到架构和学习算法的选择。我们将一起称这些选择为一个模型。对于多模态数据，已经提出了大量不同的深度学习模型。虽然有几种方法可以对它们进行划分和组织以进行审查，但我们选择根据它们的学习范式对显著的例子进行分组，特别是它们是生成的、鉴别性的还是混合模型。我们选择这种分类的原因是，这种选择会影响要选择的可用的架构和学习算法。

生成模型隐式或显式地表示数据分布，通常允许通过进程采样或“生成”新数据，因此它们的名称。另一方面，歧视性的模式则不那么雄心勃勃了。它们试图建模分布，而不是建模类边界。在监督学习设置中，我们有数据X和标签Y，生成模型学习联合概率P(X，Y)。相比之下，判别模型主要用于预测任务，这些模型学习条件P(Y|X)。然而，生成模型仍然可以具有鉴别特性。生成模型的一个优点是它们更加灵活。例如，P(X，Y)可以在推理过程中缺少模式的情况下进行采样。

鉴别模型

区分性深度体系结构直接对从输入到输出的映射进行建模，并学习模型参数

表3。多模态深度学习的不同应用。

参考资料	年数年	操作模式	出现的问题	融合的方法	模型	架构
Ngiam等人。[6]	2011	音频、视频	语音分类	中间体	生成的	稀疏的RBM
斯里瓦斯塔瓦和萨拉克特迪诺夫[30]	2012	图像、文本	图像标注	中间体	生成的	Dbn
曹操等人。[31]	2014	医学图像，文本描述	基于内容的医学图像检索	中间体	生成的	Dbm
梁等人。[32]	2015	基因表达、DNA甲基化和药物反应	癌症亚型聚集	中间体	生成的	Dbm
瓦拉达etal。[15]	2016	多光谱图像	语义分割	早期的时间	歧视性的	Fcnn
西蒙尼扬和齐瑟曼[33]	2014	图像和光流	行动识别	迟到了	歧视性的	CNN电视频道
Kahou等人。[11]	2015	视频、音频	情绪识别	迟到了	歧视性的	cnn、rnn、svm、和AE
刘等人。[20]	2015	磁磁I，EI	医学诊断	中间体	歧视性的	堆叠AE，SVM
茯苓等人。[34]	2015	视频、音频、文本	情感分析	中间期、晚一期	歧视性的	Cnn，svm
Lenz等人。[28]	2015	强度，深度视频	机器人技术的抓取	中间体	歧视性的	堆叠的AE和MLP
耆那教等人。[35]	2016	视频功能，GPS坐标，车辆动态	驾驶员对驾驶活动的预期	中间体	歧视性的	Lstm

通过最小化一些正则化的损失函数。这些模型构成了多模态学习的主要模型，而任务包括对各种问题领域的分类或识别。

除了上述积极的研究问题外，图像字幕和VQA[36]都结合了自然语言处理和机器学习算法的高级场景解释，也引起了积极的研究兴趣。在深度图像字幕中，该模型需要生成图像内容的文本描述，这可以通过使用判别技术[37]、[38]和生成方法[39]来实现。另一方面，VQA通常要求模型基于图像内容来回答复杂的问题，这是一项生成任务。这个问题也可以转换为一个鉴别设置（例如，多项选择题）[40]。最近，Kim等人。[41]为一个多模态VQA问题扩展了高度成功的深度残余网络模型。由于多种模式可能有相关性，作者精心设计了跨模式的联合残余映射，并实现了最先进的VQA结果。

鉴别性深度多模态学习模型也被提出用于人类活动识别。随着rgb深度(RGB-D)相机的廉价可用性，以及具有众多传感器的智能手机的普及，涉及4到五种模式的深度多模式学习架构已经被报道。这些问题涉及到时间数据（视频、联合运动、音频），因此，有效地学习时空依赖性是至关重要的。为了捕获时间结构和关系，深度多模态学习方法通常使用时间组件，如lstm或隐藏的马尔可夫模型，并结合视觉代表-

重压学习层，如cnn或3-d-CNNs[42]，[43]。这些模型受益于cnn和循环层的组合，它们可以共同捕获时空关系。

也有一些工作实例中，生成模型已被用于执行鉴别任务。例如，拉罗切尔和Bengio[44]提出了RBM[深度信念网络(dbn)和深度玻尔兹曼机器的构建块]的一个判别变体。在“人类活动识别”、“医学应用”和“自主系统”章节中讨论应用领域时，已经提到了其他判别模型。此外，表3还列出了其他多模态问题的判别模型的例子。虽然鉴别模型在分类或回归任务方面表现出色，但当缺少数据或模式时，它们无法处理。判别模型还需要大量标记数据，在某些应用中获得可能很昂贵。接下来，我们回顾了深度生成多模态模型，考虑到学习多模态表示的缺点，在学习多模态表示的背景下提供了一些优势。

生成的模型

深度生成模型通常表征观察或可见数据的高阶相关特性，用于模式分析或综合目的。它们还可以用来表征可见数据及其相关类的联合统计分布。像dbn这样的生成模型也可以用于分类和回归任务，利用它们从无标记数据中学习（无监督），并在区分设置中进行微调

反向传播算法或通过使用学习表示与其他分类器，如支持向量机（支持向量机）。

对于多模态学习问题，生成模型在测试期间可能缺少模式或缺乏标记数据的情况下是有用的。Ngiam等人的早期作品。[6]、斯里瓦斯塔瓦和萨拉克特迪诺夫[30]证明了生成模型确实能够处理这类学习问题。从那时起，文献中报道了一些专门处理在缺少数据[31]，[45]的情况下使用生成式深度多模态网络的工作。

虽然基于堆叠rbm的基于能量的模型在深度生成多模态学习中受到了大部分的关注，但生成模型的格局正在发生变化。最近，生成对抗网络[46]，用变分推理[47]训练的深度定向模型，在多模态和单模态设置[48]-[50]中获得了吸引力。

混合动力汽车的模型

当有别模型被训练以最大限度地提高类之间的分离时，生成模型擅长于建模数据分布。混合模型在一个统一的框架中结合了区分组件和生成组件。邓
[51]将混合深度体系结构定义为目标是识别，但通过辅助（通常以一种重要的方式）获得生成架构结果的体系结构。例如，混合模型中的生成组件可以学习输入模式的深度表示，并使用判别组件进行分类或回归任务。

混合模型可按[52]分为三组：

1) 优化单个目标函数以使用生成和鉴别组件学习联合表示的联合方法
2) 迭代方法学习共享表示使用迭代方法，如使用从生成和鉴别组件更新的表示的期望最大化
3) 分阶段的方法，其中生成和鉴别组件分别分阶段训练。

生成模型以无监督的方式学习的表示可以使用监督训练作为鉴别成分的特征。

联合模型的一个例子报道在[53]，其中短期时间特征和长期时间依赖音频视频模式建模通过结合条件RBM时间生成模型为前者和鉴别组件组成的条件随机场为后者。该模型还能够处理由于生成组件而导致的缺失模式。其他相关的混合架构包括Sachan等人。[54]和Liu等人。[55].

总结

在本节中，我们将根据它们的主要学习范式来强调多模式架构。在某种意义上说，深度学习模型可以被认为是构建块，允许我们“混合和匹配”不同的模型，以创建复杂的深度多模态架构。虽然这可以被视为深度学习的一个优势，但一个常见的问题是，架构设计更像是一门艺术，而不是一门科学。尽管如此，与每个模型都有许多关联的超参数必须仔细微调，在处理混合架构时，这个过程可能会更加复杂。另一个需要关注的方面是模式及其表示之间的融合结构的选择。接下来，我们将讨论多模态融合结构的几种选择，并将我们的讨论引导到优化和学习该融合架构以提高性能的有吸引力的概念上。

熔变结构

深度架构提供了在早期、中期或晚期融合时实现多模态融合的灵活性。在深度学习出现之前的多模态融合方法通常将早期融合称为特征级融合，将晚期融合称为决策级融合。然而，通过深度学习方法，特征级融合的想法可以进一步扩展到中间融合的概念。

早期的融合

早期的融合包括将多个数据源，有时非常不同，集成到单个特征向量中，然后被用作机器学习算法的输入，如图2(a)所示。要融合的数据是来自传感器的原始或预处理数据；因此，经常使用术语数据融合或多感觉器融合。

如果不需要特征提取就可以进行数据融合，这可能是相当具有挑战性的。例如，不同传感器之间的采样率可能会有所不同，或者如果一个源产生离散的数据，而另一个源提供连续的数据流，则来自多个数据源的同步数据可能不可用。

为了缓解一些与融合原始数据相关的问题，我们首先可以从每个模态中提取高级表示，这可以是手工制作的特征，也可以是学习表示，这是在深度学习中常见的那样。当使用非层次特征时，就像手工制作的特征一样，在从输入机器学习算法之前，从多个层次中提取的特征只能在一个层次上融合。由于深度学习本质上涉及到从原始数据中学习层次表示，这就导致了中级融合。

大多数早期融合模型都简化了一个假设，即不同信息源的状态之间存在条件独立性。这在实践中可能不是正确的，因为多种模式往往是高度相关的（例如，视频和深度线索）。Sebe[56]认为，不同的流包含仅在高级别的级别与另一个流相关的信息。在[57]中可以看到一个很好的例子。这个假设允许每个模态的输出独立于其他模态进行处理。

在其最简单的形式中，早期融合涉及到茯苓等人所实现的多模态特征的连接。[34]. 多模态数据的早期融合可能不能充分利用所涉及的模态的互补性，并可能导致可能包含冗余的非常大的输入向量。通常，应用像PCA这样的降维技术来消除输入空间中的这些冗余。自动编码器是PCA[58]的非线性推广，被广泛用于深度学习，从原始数据中提取分布式表示。这一想法已被扩展到学习多模态嵌入空间，目的是在共同特征空间[59]，[60]中表示多模态数据。

在多模态数据的早期融合中所面临的问题之一是确定不同数据源之间的时间同步性。通常，这些信号以一个共同的采样速率重新采样。为了缓解这一问题，马丁内斯和扬纳卡基斯·[61]提出了几种方法（卷积、训练和池化融合）来整合离散事件序列与连续信号。

晚期核聚变

Lateor决策级融合是指来自多个分类器的决策的聚合，每个分类器都以不同的模式进行训练[见图2(b)]。这种融合架构往往更受青睐，因为来自多个分类器的错误往往不相关，而且该方法是特征无关的。存在不同的规则来决定如何组合来自不同分类器的决策。

这些融合规则可以是最大融合、平均融合、基于贝叶斯规则的，甚至可以使用元分类器学习。决策级融合在21世纪中期就很流行了，当时集成分类器在机器学习社区中受到了广泛的兴趣。

除了表3中列出的一些工作外，已经有一些工作对深度多模态[33]、[43]、[62]学习使用了晚期或决策级融合。基于我们所回顾的论文，我们没有发现晚期融合比早期融合更好的决定性证据——性能在很大程度上依赖于问题。毫无疑问，当输入模式显著不相关，具有非常不同的维度和采样率，对于多模态学习问题实现后期融合方法要简单得多。另一种方法，中间融合，为如何以及何时可以融合从多模态数据中学习到的表示提供了更多的灵活性。

中间的核聚变

神经网络通过一层管道映射输入，将原始输入转换为更高层次的表示。每一层通常交替进行线性和非线性操作，以缩放、位移和倾斜其输入，产生原始数据的新表示。在多模态上下文中，当所有的模态都被转换为表示时，它就可以将不同的表示融合到单个隐藏层中，然后学习一个联合的多模态表示。深度多模态融合的大多数工作都采用了这种中间融合方法，其中共享表示层是通过合并来自多个模态特定路径进入该层的连接来构建共享表示层。图2(c)展示了一个简单的具有三种模式的中间融合模型。表示（特征）使用不同类型的层(如二维卷积、三维卷积或完全连接）进行学习，表示使用融合层进行融合，也称为共享表示层。

这个共享表示层可以是一个单个共享层，它在某个深度融合多个通道，也可以逐渐融合，一次融合一个或多个模式。共享表示层中特征或权值的初始连接可能导致过拟合或网络由于不同的底层分布而无法学习模式之间的关联。提高多模态融合性能的一种简单方法是，在构建简单的共享表示层（或融合层）后，应用某种形式的降维方法，如PCA[63]或堆叠自动编码器[10]。与其他融合技术相比，这种在不同深度融合的各种表示形式的选择可能是深度多模态融合中最强大和最灵活的方面。柔性核聚变方案的优点可以在

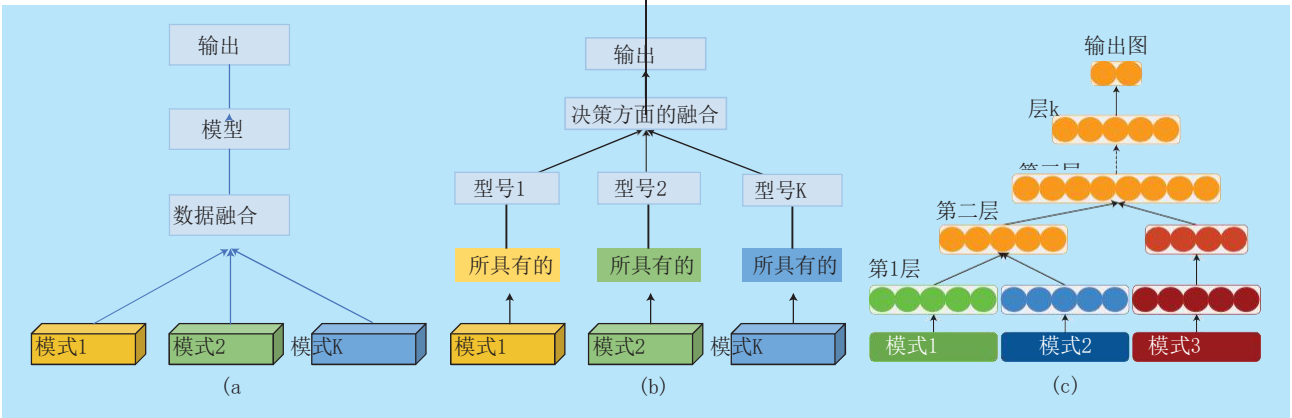


图2。为多模态学习的各种融合模型的说明。(a)早期或数据级融合，(b)晚期或决策级融合，以及(c)中间融合。

卡尔帕蒂等人的工作。[64]表明，使用“慢融合”模型，学习到的视频流表示在训练过程中逐渐融合，与早期融合和晚期融合模型相比，大规模视频分类问题始终产生更好的结果。同样地，Neverova等人。[8]的经验表明，通过实施逐步融合策略，首先将高度相关的方式融合到较少相关的方式（例如，视觉方式首先，然后是运动捕捉，然后是音频），产生了交际手势识别的最先进结果。

尽管使用共享表示层学习多模态表示确实是灵活的，但许多当前的架构需要仔细设计如何、何时以及可以融合哪种模式。在“融合结构学习和优化”部分中，我们将讨论对优化多模态学习所需的繁琐架构设计过程的进一步尝试。

多模态正则化化

深度学习技术通过最小化一个损失函数，迭代地优化一组模型参数（通常是每一层之间的权重和偏差）。为了改进泛化，使用了一个或多个正则化策略，通常作为损失函数的附加项。从计算的角度来看，正则化为导致算法加速的优化问题提供了稳定性，并且，从统计的角度来看，正则化减少了过拟合的[65]。

在深度多模态学习环境中，一个重要的设计考虑是成本函数和正则化器的制定，它们执行模式间和内态关系，如信息论正则化器和结构化正则化，我们现在简要回顾这些。

信息理论的正则化器是利用互信息和信息的变化等度量来制定的。例如，Sohn等人。[66]提出了一个成本函数，最小化模式之间的信息变化，以学习模式之间的关系。这个公式背后的直觉是，学习最大限度地提高一种数据模式关于其他数据模式的信息量，将允许生成模型解释给定部分观察的缺失数据模式。或者，一个互信息项也可以在训练[67]中最大化。Zhu等人提出了另一种基于Kulback-Leibler(KL)发散的信息理论损失公式。[68]中的多标签图像注释问题。在训练前阶段，他们首先训练cnn使用未标记数据从每个模态中学习中间表示，然后，在微调阶段，使用反向传播来最小化预测分布和地面真实分布之间的kl差异。最后，为了学习多模态权值的最优组合，他们采用指数化的在线学习算法，依次寻找最优的组合权值集。

Wu等人从多任务学习[69]中的结构化特征选择为灵感。[70]设计了一个模型使用一个跟踪范数正则化术语，它鼓励相似的模式使用视频和音频模式为视频分类问题共享相似的表示。

Wang等人也探索了强制相互间和相互内相关性的成本函数。[71]. 他们的公式包括一个判别项，和一个基于正则相关分析的相关项。在后续的工作[72]中，他们提出了一个多模态融合层，该层使用矩阵转换显式地强制执行由不同模态特征共享的公共部分，同时保留特定于模态的学习。

Lenz等人。[28]在成本函数中制定了一个结构化的正则化术语，它允许模型学习多个输入模式之间的相关特征，但规范每个特征使用的模态数量，从而阻止模型学习模式之间的弱相关性。结构化正则化本质上对每组特定的权值分别应用于某种形式的正则化。他们考虑了一个多模态机器人抓取的结构正则化的几种变体要执行的任务。在他们的案例中效果良好的一个合并了L在最大标准惩罚之上的0标准

$$\mathrm{f(W)}=//\mathrm{I}_{j=1r=1}^{K}\text{最大}S_{ri}\$\; iW_{i,1},\; |j>0|_{j}$$

其中，S_{ri}如果特征i属于r组，并且是另一个-明智的零。S是一个大小为r#N的二进制模态矩阵，其中每个元素S_{ri}表示一个可见单元的成员身份，

x_i，在一个特定的模式中，r。我是一个指示器函数如果它的参数为true，否则为零，则取一个值为1。

在一些问题中，时间上下文可以发挥重要的作用，例如，驾驶员活动预期。与人类活动识别不同，即完整的时间上下文可用，在驾驶员活动预期中，机器学习系统必须在事件发生前的短时间内只使用部分上下文进行预测。为了解决这个问题，耆那教等人。[35]将一个随时间呈指数增长的时间项纳入到具有LSTM单元的多模态RNN的成本函数中。这鼓励了模型尽早修复错误。

多模态感知的正则化器已经导致了模型性能的边际到显著的改进。尽管包括了这些多模态正则化策略，但本节中讨论的深度学习架构还是将输入模式合并到单个融合层中。一个可能的扩展可能是研究一个利用这些正则化策略的渐进融合模型。

融合结构的学习与优化

迄今为止提出的大多数多模态深度学习架构都是精心手工制作的。虽然许多模型采用了一个单一的融合层（共享表示层），但一些突出的工作[8]，[64]实现了一个渐进的融合策略。融合方式的选择

表示的深度，通常是基于直觉（例如，早期融合相似的模式，然后在更深的层次上融合不同的模式）。当涉及到两种以上的模式时，这也取决于问题中所使用的模式的性质，选择一个最佳的融合架构可能更具挑战性。一个自然的进展将是通过将其转换为一个模型搜索或结构学习问题来寻找一个最优的多模态融合架构。

单峰问题的神经网络结构优化一直被机器学习研究者研究。这些主要涉及到确定一个网络中的最佳神经元数量和层数。在网络良好的泛化能力和参数的数量和训练数据的可用性之间存在着权衡。太大的网络可能会表现得良好或过拟合，这取决于它是否使用足够大的训练数据进行训练，而太小的网络，可能会拟合不足，并可能导致泛化性较差。

一种常见的方法是采用一种自下而上的建设性的方法。Elman[73]提出的基本思想是从一个相对较小的网络开始，并逐步添加隐藏的单元或图层，直到找到性能最好的体系结构。最近，在大规模的环境下，陈等人。[74]通过在一个神经网络之间到另一个神经网络之间的知识转移，逐渐为初始风格的[75]网络增加了深度和宽度。

修剪算法[76]从自上而下的方法解决了同样的问题。最近针对DNNs的方法包括Feng和Darrel[77]的工作，他们提出了一种不断发展的生长和修剪算法来优化印度自助餐过程的结构——CNN模型，以及Yang等人。[78]为基于稀疏表示的大型、不同的数据集引入了网络剪枝。

基于遗传算法(GA)的神经网络结构优化是最早用于神经网络结构搜索和优化[79]的元启发式搜索算法之一。21世纪初，一种名为增强拓扑神经进化(NEAT)[80]的算法，该算法也使用GAs进化出越来越复杂的神经网络结构，受到了广泛关注。最近，真崎和渡边[81]应用GAs和协方差矩阵演化策略来优化DNN的结构，将DNN的结构参数化为基于有向无环图表示的简单二进制向量。由于GA搜索空间可能非常大，而且搜索空间中的每个模型评估都很昂贵，因此使用大型GPU集群进行并行搜索来加快搜索过程。

如果设计出一个合适的网络结构表示，并且在搜索过程中训练和测试多种架构的成本并不昂贵，那么这些神经网络结构搜索和优化技术的成本可以很容易地扩展到多模态设置。随着数据集大小接近gb，甚至是tb的级别，以及涉及数百万个参数和多种模式的深度网络架构，搜索和优化多模态融合结构可以除非实现了一些并行搜索过程或使用了有效的优化算法，否则代价非常昂贵。虽然贝叶斯优化(BO)[82]是超参数优化的热门选择，但最近已被用于多模态融合架构优化[83]。通过使用基于高斯过程的BO搜索所有可能的多模态融合架构的空间，将架构优化转换为一个离散优化问题。提出了一种新的图诱导核来量化搜索空间中不同架构之间的距离。

强化学习[84]也被用于深度神经体系结构搜索[85]。本文提出了一种利用RNN生成神经网络可变长度模型描述的新方法。RNN通过强化学习进行训练，以最大化生成架构在验证集上生成的预期精度。

最近的一些工作已经将结构学习作为网络中正则化或容量控制的一种方法。通过随机剪枝网络，随机正则化方法可以看作是一种通过模型平均来提高泛化能力的集成。库尔karni等人。[86]实现了一种通过确定性正则化来学习dnn结构的方法。它们在每对完全连接的层之间插入一个稀疏的对角线矩阵其条目是I被惩罚。这就隐式地定义了大小每一层的有效权值矩阵。该方法与辍学的[87]也有类似的效果。块[88]可以通过一种聪明的技术同时执行正则化和模型选择，将隐藏单元随机分配给“集群”，形成块结构化的权值矩阵。此外，通过平均多个随机推理通道的输出（这可以看作是集成分类器的一个例子），结果优于ResNets。该架构有效地实现了多个架构的后期融合，以获得更好的结果。

随机正则化已推广到多模态
背景：无诺娃等人。[8]和最近的Li等人。 [89]. 在后一项工作中，作者表明，当模态间相关性较高时，早期融合方法（其融合结构被网络学习）产生更好的结果，而当输入模态相关性较低时，晚期融合方法效果更好。这与前者所作出的经验证选择相一致。

在本节中，我们介绍了一些最近使用随机正则化或优化的深度多模态融合架构，其性能与精心设计的相媲美或更好。虽然特征工程已经在很大程度上通过深度表示学习来解决，但下一个逻辑步骤将是放弃深度架构的细致工程，并追求自动实现这一目标的技术。

数据集

为了促进多模态学习的研究，许多数据集已经发布给公众。我们注意到，大多数的

这些数据集通常涉及以人为中心的视觉理解，其变量包括情绪识别、群体行为分析等。表2列出了许多这样的数据集、所涉及的模式和问题领域。虽然这个列表并不详尽，但我们涵盖了可用于多模态研究的最近的数据集（其中许多是在过去三年中发布的）。虽然大多数数据集包括至少两种模式（例如，图像和文本）或最多四种(RGB-D、音频和骨骼姿势)，但一些数据集，例如，H-MOG[12]，包括最多9种不同的模式。对于感兴趣的读者，Firman[90]对102个RGB-D数据集进行了广泛的调查。自主驾驶和驾驶员辅助系统（使用驾驶员行为预测）正在作为深度学习的一个流行的研究课题。这种数据集不仅是高度多模态的[91]，数据最多来自多达6个单独的传感器，而且还有非常大的-数小时的可用数据。例如，牛津机器人车[92]数据集包含在不同天气条件下超过23tb的全年驾驶数据。

我们注意到，可用的多模式医疗数据集相对较少，可能是由于成本、伦理和隐私问题。大多数医疗数据集也往往要小得多，涉及10到50名受试者，也遭受阶级失衡（例如，与异常病例相比，正常病例更常见）。医学信息学和成像研究严重依赖于多模态信息，这可以用于改善计算机辅助诊断。我们鼓励人们努力收集和公开此类数据集。

研究结论及未来的发展方向

在本文中，我们回顾了深度多模态学习的最新进展。不可否认的是，在学习问题中合并多种模式几乎总是会对各种问题产生更好的性能。从融合的角度来看，我们看到深度多模态学习技术可以分为听觉和晚期融合方法，深度学习方法促进了灵活的中间融合方法，这不仅简化了融合模态表示和学习联合表示，而且允许多模态融合。虽然在许多情况下，深度学习减少了对特征工程的需求，但深度学习架构仍然涉及大量的手工设计，而且实验者可能没有探索可能的融合架构的全部空间。研究人员应该将学习的概念扩展到体系结构，以努力获得一种真正通用的学习方法，该方法可以在最少或没有人工干预的情况下适应于特定的任务。

我们回顾了学习最佳架构的几个选项。这包括随机正则化，将架构优化转换为一个超参数优化问题，例如，使用BO，和增量在线强化学习。在我们看来，这是深度多模态学习中最令人兴奋的研究领域。体系结构学习可能非常计算密集型，因此研究人员应该利用硬件加速和分布式深度学习的进步。

我们还确定了几个在深度多模态学习中获得最多关注的应用领域。这包括RGB-D和来自手机上的许多传感器的数据，这些数据已被用于涉及多模态数据的一系列问题，如人类活动识别及其变体。我们预计，这一领域将在未来几年获得更多的关注，这将深刻地影响我们的日常生活。另一个突出的重要领域是医学研究，它涉及许多数据模式，其中一些如果没有人类专家，就很难解释。随着医学界开放人工智能辅助诊断的兴起，我们将看到这一领域取得更重要的进展。最后，另外两个正在获得深度学习研究人员注意的应用程序领域包括自动驾驶汽车或机器人技术和多媒体应用程序，例如，视频转录、图像字幕等。新的应用程序，如使用多模态输入的在线聊天机器人，如图像，以及使用多模态数据的文本或推荐系统，可能会在不久的将来得到广泛应用。

最后，我们承认这是一个非常快速发展的领域，而且随着新研究发表的速度，深度多模态学习体系结构中的许多新创新必然会被提出。我们试图不为体系结构设计提供具体的建议，因为我们发现许多问题需要应用程序特定的考虑。无论如何，我们认为这是一个及时的出版物，作为我们所强调的未来研究的方向，希望能够作为更有组织的努力推进研究领域的指导。

的作者

拉马钱德拉姆(dramacha@uoguelph.ca)收到了他的作品B. 工业技术技术学位及博士学位。D.他分别于1997年和2003年在马来西亚圣大学获得计算机视觉和机器人学学位，并曾担任副教授。他是加拿大安大略省圭尔夫大学的研究员，也是IEEE的高级成员。他对计算机视觉、医学成像和多模态问题等方面的深度学习问题很感兴趣。

格雷厄姆·W。 泰勒(gwtaylor@uoguelph.ca)分别于2003年和2004年在加拿大滑铁卢大学获得了应用科学的学士和硕士学位。他获得了博士学位。D.2009年在加拿大多伦多大学获得计算机科学学位，他的论文联合顾问是杰弗里·辛顿和山姆·罗威斯。他是加拿大安大略省圭尔夫大学的副教授，人工智能向量研究所的成员，以及加拿大高级研究所的阿兹里利全球学者。他对统计机器学习和生物启发的计算机视觉很感兴趣，重点是无监督学习和时间序列分析。

参考文献

[1] Y.莱昂村，Y。《深度学习》，《自然》，第1卷。521年，没有。 7553,每页。436－444,2015.

[2] D. 拉哈特，T。Adali和C.Jutten，“多模态数据融合：方法、挑战和前景的概述”，序言。IEEE，第1卷。103年，没有。第9页。1449－1477,2015.

[3] P.K.阿特里伊，米1.A.侯赛因，《多媒体分析的多模态融合：调查》，多媒体系统，第1卷。16岁、没有。第6页。 345－ 379, 2010.

[4] B. 哈勒吉，A。卡米斯，F。O.Karray和S.N.Razavi，“多传感器数据融合：最先进的回顾”。*融合版，第1卷*。14岁、没有。第1页。28－44, 2013.

[5] *L.I.Kuncheva，结合模式分类器：方法和算法。霍博肯，新泽西州：Wiley，2004*。.

[6] J.恩吉亚姆，A.科斯拉，M.金，J。南，H.李，和A。Y.Ng，“多模式深度学习”，在Proc中。*第28英寸。续称。机器学习（ICML-11），2011年，第页*。 689－696.

[7] N.斯里瓦斯塔瓦和R.R.。*萨拉克特蒂诺夫，“使用深度波尔兹曼机器的多模态学习”，在Proc。神经信息学的研究进展。处理系统系统。*，2012年，第2页。2222－ 2230.

[8] N.无娃，C。狼，G。泰勒，和F。网络，“ModDroop：自适应多模态手势识别”，IEEE跨。*模式的肛门。马赫数。被告知。*，第1卷。38岁、没有。第8页。 1692－ 1706, 2016.

[9] S.S.墨克吉和N。M.罗伯逊，“深头姿态：多模态视频中的凝视方向估计”，IEEE反式。*多媒体，第1卷*。17岁、没有。第11页。 2094－2107, 2015.

[10] C.丁丁和D。陶，“通过多模态深度人脸表示的稳健人脸识别”，IEEE反式。*多媒体，第1卷*。17岁、没有。第11页。 2049－2058, 2015.

[11] S. E.卡侯，X号。布思利尔，P。兰布林，C。古尔塞雷，五世。米卡尔斯基，K。康达，S。让，P。弗鲁门蒂等人，“情感网络：视频中情感识别的多模式深度学习方法”，J.多媒体用户界面，第1卷。10岁、没有。第2页。99－111, 2015.

[12] Z.西托娃，J。Sedeˇnka，Q.杨，G。彭，周鹏。加斯蒂和巴拉加尼，“HMOG：智能手机用户持续认证的新行为生物特征特征”，IEEE变性。*国际。取证安全部，第1卷*。11岁、没有。第5页。 877－892, 2016.

[13] V.拉德，N。 D.巷，S。巴塔查里亚，C。马斯科洛先生。K.码头，和F。Kawsar，“面向移动设备上的多模式深度学习的活动识别”，在Proc。*ACM内部信息。联合连接。普适和无处不在的计算：辅助计算，2016年，第页*。185－ 188.

[14] A. 耆那教，J.汤普森，Y。LeCun，和C。“莫迪普：一个使用运动特征进行人类姿态估计的深度学习框架”，在Proc中。*亚洲国家。美国计算机视觉出版社，2014年，第3页*。 302－315.

[15] A.瓦拉达公司。奥利维拉，T。布罗克斯，和W。伯加德，“使用多模态融合对森林环境的深度多光谱语义场景理解”，在Proc。内部信息。*符号。《实验机器人学》(ISER，2016年)，2016年，第3页*。 465－477.

[16] D.沈先生，G。吴，和h.i。苏克，“医学图像分析中的深度学习”，安努。*回顾一下生物医学工程公司。*，第1卷。第19页。 221－248, 2017.

[17] R.Kiros，K.波波里，博士。科布萨斯，和M。“领域独立医学图像分割的叠加多尺度特征学习”，在Proc。内部信息。关于马赫数的研讨会。*医学影像学学习，2014年，第页*。 25－32.

[18] P.吴先生，南卡罗来纳州。夏和。赵先生，D.王，和 C. 苗，“在线多模态深度相似性学习与在图像检索中的应用”，在Proc。*第21个ACM英寸。续称。多媒体出版社，2013年，第3页*。. 153－162.

[19] M.西莫诺夫斯基，古铁雷斯-贝克尔，D.马特乌斯，N。纳瓦布，和N。科莫达基斯，“多模态注册的深度度量”，在Proc。内部信息。续称。医学图像计算机与计算机辅助干预，2016年，第页。 10－18.

[20] S.Liu，S.Liu，W。蔡，蔡，蔡，蔡。冯，医学。Fulham等人，“多类诊断阿尔茨海默病的多模式神经成像神经系统学习”，IEEE跨式。*弯曲的材料。注册工程师*。，第1卷。62岁、没有。第4页。1132－1140, 2015.

[21] P.变形虫，维埃拉，普京，和 A. ，“深度学习在生物医学中的应用”，分子制药学，第1卷。13岁、没有。第5页。 1445－1454, 2016.

[22] Y.郭，G。吴先生，洛杉矶。指挥官，S。大小为，五号。朱厄尔斯，W。林，和D。*沈，“通过与退化特征匹配的稀疏补丁从婴儿大脑中分割海马体”，在Proc中。内部信息。续称。医学图像计算机和计算机辅助干预，2014，p*。 308.

[23] N.塔杰巴克什，J。Y.小腿，S。R.古鲁都，R。T.赫斯特，C。 B.肯德尔，M。 B.Gotway，和J。梁，“医学图像分析的卷积神经网络：全面训练或微调？”” IEEE变性。*地中海的人。伊玛目集团*。，第1卷。35岁、没有。第5页。 1299－1312, 2016.

[24] I.Sturm，S.拉普斯金，W。*和r。“单次试验eeg分类的可解释深度神经网络”，J.神经科学方法，第1卷*。第274页。12月141日至1415日。 2016.

[25] S.G.Kim，N。他们的时代方先生，哈万尼先生，A.格拉玛，和S。“打开黑盒：一个可解释的基于深度神经网络的分类器，用于细胞类型特定的增强器预测，”BMCSyst。生物学》，第1卷。10岁、没有。2，p.54, 2016.

[26] C.陈，A.。塞夫，A。科恩豪泽，和J。肖，“深度驾驶：学习提供自动驾驶在直接感知”。IEEE国际公司。续称。美国计算机视觉出版社，2015年，第3页。 2722－2730.

[27] J.长，E。谢尔哈默，和T。达雷尔，“语义分割的全卷积网络”，在Proc中。*IEEE的续称。计算机视觉与模式识别，2015年，第3页。* 3431－3440.

[28] I.Lenz，H.Lee，和A.Saxena，“检测机器人抓取的深度学习”，Int。*J.机器人学的经验。*，第1卷。34岁、没有。第4-5页。 705－724, 2015.

[29] S.顾uE。霍莉，T。非法说唱乐，和莱文。(2016). 使用异步非策略更新对机器人操作的深度强化学习。*arXiv。*[在线的]。
可用的是：https://arxiv.org/abs/1610.00633

[30] *N.斯里瓦斯塔瓦和R.萨拉克胡特蒂诺夫，“学习具有深度信念网的多模态数据的表示”，在Proc上提出。第29英寸。续称。机器学习（研讨会），2012年。*

[31] Y.曹操，S。斯蒂菲，J。他，D。肖，C。陶先生，P。陈，和H。穆勒，“医学图像检索：多模态方法”，癌症信息学，第1卷。13岁、没有。第三期，第页。 125, 2014.

[32] M.梁，李，T。陈，和J。曾，“采用多模式深度学习方法对多平台癌症数据的综合数据分析”，IEEE/ACM跨。*组合件。生物醇。生物蛋白f。*，第1卷。12岁、没有。第4页。 928－937, 2015.

[33] K.齐瑟曼，“视频中动作识别的双流卷积网络”。*《神经信息处理系统的研究进展》，2014年，第3页。* 568－576.

[34] S.茯苓，E。坎布里亚岛，和A岛。*Gelbukh，“深度卷积神经网络文本特征和多核学习的话语级多模态情绪分析”，在Proc。续称。《自然语言处理的经验方法》，2015年，第3页。2539－2544.*

[35] A. 耆那教，辛格。，“通过感觉融合结构进行驾驶员活动预期的递归神经网络”，在Proc中。*2016年IEEEInt。续称。机器人与自动化(ICRA)，2016年，第页。* 3118－3125.

[36] S.安托尔，A。阿格拉瓦尔，J。陆氏市，M.。米切尔，美国博士。巴特拉，C。L.齐特尼克，和D。Parikh，“VQA：视觉问题回答”，在Proc。*内部信息。续称。计算机视觉(ICCV)，2015年，第页。* 2425－2433.

[37] J.多纳休市，L。安妮·亨德里克斯，S。瓜达拉马先生。罗赫巴赫，S。维毒兰，
K.申科，和T。*达雷尔，“视觉识别和描述的长期循环卷积网络”，在Proc。IEEE的续称。计算机视觉与模式识别，2015年，第3页。2625－2634.*

[38] A. 卡帕蒂，朱林和F.F.F。李，“双向图像句子映射的深度片段嵌入”，在Proc中。*《神经信息处理系统的研究进展》，2014年，第3页。* 1889－1897.

[39] O.葡萄酒，A。托谢夫，S。本吉奥，和D。*厄尔汉恩，“显示和告诉：一个神经图像字幕发生器，在Proc”。IEEE的续称。计算机视觉与模式识别，2015年，第3页。* 3156－3164.

[40] M.*任，R.基罗斯，和R.Zemel，“探索模型和数据的图像问题回答，”，在序言中。《神经信息处理系统的研究进展》，2015年，第3页。2953－2961.*

[41] J.-H.金姆，西南部。李，h博士。夸克，行政硕士。熙熙，J。金，j.w。哈，和b.t。张先生。(2016). 视觉QA的多模态残差学习。. ArXiv[在线]。可用的：https://arxiv.org/abs/1606.01455

[42] F.J.奥尔多内斯和D。Roggen，“多模态可穿戴活动识别的深度卷积和LSTM递归神经网络”，传感器，卷。16岁、没有。1，p.115, 2016.

[43] D.吴，L。庇沟，j。金德曼斯，N。D.-H.勒，L.。邵氏星，J。丹布雷，和j.m。*Odobez，“用于多模态手势分割和识别的深度动态神经网络”，IEEE反式。模式的肛门。马赫数。被告知。*，第1卷。38岁、没有。第8页。1583－1597, 2016.

[44] H.拉罗切尔和Y。本吉奥，“使用判别限制玻尔兹曼机器的分类”，在Proc。*第25英寸。续称。机器学习，2008年，第3页。* 536－543.

[45] Y.黄，W。王，和L。王，“无约束的多模态多标签学习”，IEEE变性。*多媒体，第1卷。*17岁、没有。第11页。 1923－1935, 2015.

[46] I.古德费罗，小普吉特-阿巴迪，米尔扎先生，B.徐沃德法利，奥泽尔，考维尔和Y。*本加奥，“生成的对抗网”，在Proc。《神经信息处理系统的研究进展》，2014年，第3页。2672－2680.*

[47] D.P.金玛和M。焊接，“自动编码变分贝叶斯”。内部信息。*续称。《学习代表》(ICLR)，2014年。*

[48] S.Reed，Z.Akata，X。严，李，席勒，和李，“图像合成的生成对抗文本”，序言。第33英寸。续称。机器学习(ICML)，2016年，pp页。1060－1069.

[49] 铃木M.，中山和Y。松夫。(2016). 将多模态学习与深度生成模型联合学习。. ArXiv[在线]。可用的：https://arxiv.org/abs/1611.01891

[50] G.潘迪和A。杜克基帕蒂。（2016）。条件多模态深度学习的变分方法。*arXiv。*[在线的]。可用的是：https://arxiv.org/abs/1603.01801

[51] L.邓，“一个关于深度学习的架构、算法和应用程序的教程调查”，APSIPATrans。*信号和信息。处理过程，第1卷。*第3页。 1－29, 2014.

[52] M.R.Amer，T。护板，B.西迪基，A.塔姆拉卡尔，“深度多模式融合：混合方法”，国际大学。j。视觉，页。1－17, 2017. 多：10.1007/s11263-017-0997-7。

[53] M.R.电流，B。西迪基，美国标准大学。可汗，A。迪瓦卡兰，和H。索威尼，“使用动态混合模型的多模态混合”，在Proc中。IEEE2014在计算机视觉冬季领域的应用。，2014年，第3页。556－563.

[54] D.S.Sachan，美国。特克瓦尼和A.塞蒂，“使用深度神经网络从多模态信息的运动视频分类”，在Proc。*2013年人工智能进步协会秋季研讨会。*，2013年，第3页。 102－107.

[55] Y.刘先生，X.。冯和周泽伟，“具有堆叠收缩自动编码器的多模态视频分类”，信号处理，第1卷。第120页。761-766年，3月。 2016.

[56] N.Sebe，计算机视觉中的机器学习，第1卷。29.荷兰，多尔德雷希特：施普林格，2005年。

[57] A.欧文斯，J。吴先生，J。H.麦克德莫特，W。T.弗里曼，和A。托拉尔巴，环境的声音为视觉学习提供监督。章，瑞士：施普林格国际出版，2016，页。 801－816.

[58] G. E.Hinton和R.R.。萨拉克特蒂诺夫，“用神经网络减少数据的维数”，科学，第1卷。313年、没有。5786年，页。 504－507, 2006.

[59] D. 王先生，P。崔，欧，W。Zhu，“使用正交深度结构学习多模态表示的紧凑哈希代码”，IEEE反式。*多媒体，第1卷。*17岁、没有。第9页。1404－1416, 2015.

[60] J.马西，医学硕士。布朗斯坦，上午。布朗斯坦，和J。施米杜伯，“多模态保持相似性的哈希”，IEEE反式。*模式的肛门。马赫数。被告知。*，第1卷。36岁、没有。第4页。824－830, 2014.

[61] H.P.马丁内斯和G。N.扬纳卡基斯，“深度多模态融合”，在Proc。第16英寸。*续称。《多模态交互作用》，2014年，第3页。* 34－41.

[62] S. E.卡侯，C。爸爸，X岁。布思利尔，P。泡沫膜，C。古尔塞雷，R。备忘录，P。文森特，A.考尔维尔，等人，“结合模态特定的深度神经网络的视频情感识别，”，在Proc。*第15个ACM英寸。续称。多模态交互作用，2013年，第3页。* 543－550.

[63] D.李、雷、李，“异质人脸识别的共享表征学习”，在Proc中。*自动人脸和手势识别第11个IEEEInt。续称。研讨会，2015年，第3页。* 1－7.

[64] A. 卡尔帕蒂，G。托德里奇，S。谢蒂，T。梁振英和L。费飞，“带有卷积神经网络的大规模视频分类”，在Proc中。IEEE的续称。*计算机视觉与模式识别，2014年，第3页。* 1725－1732.

[65] M.J.温赖特，“高维问题的结构化正则化器：统计和计算问题”，Annu。牧师。统计数据应用程序，第1卷。第1页。4月233-253日。 2014.

[66] K.Sohn，W。*商和李志，“改进了信息变化的多模态深度学习”，在Proc。神经信息处理系统的研究进展。*，2014年，第3页。 2141－2149.

[67] J.J.-Y.王，Y。*王，赵和高，“最大互信息正则化分类”，英文。应用程序。人工告知。*，第1卷。第37页。1月1日至8日。2015.

[68] S.朱，X。李，和S。沈，“基于多模态深度网络学习的图像注释”，IET电子。*莱特的名字。*，第1卷。51岁、没有。第12页。 905－906, 2015.

[69] H.Fei和J。桓，《多任务学习的结构化特征选择与任务关系推理》，《知识与信息》。系统系统。，第1卷。35岁、没有。第2页。345－364,2013.

[70] Z.吴，y.g.。姜，J.王，J。Pu，和X。薛，“探索与深度神经网络的视频分类”，在Proc。*ACM内部信息。续称。多媒体出版社，2014年，第3页。*167－176.

[71] A.王志，J。路，J。蔡先生，T.。J.第二章，和G。Wang，“RGB-D对象识别的大边缘多模态深度学习”，IEEE反式。*多媒体，第1卷。*17岁、没有。第11页。1887-1898年，11月。 2015.

[72] A.王志，J。蔡先生，J。陆，和t.j。*Cham，“MMSS：RGB-D对象识别的多模式可共享和特定的特征学习”，在Proc中。IEEE国际公司。续称。美国计算机视觉出版社，2015年，第3页。* 1125－1133.

[73] J.“神经网络的学习和发展：开始小的重要性”，认知，卷。48岁、没有。第1页。 71－99, 1993.

[74] T.陈，我。古德费罗，和J。镜头。(2015). 网络：通过知识转移加速学习。. ArXiv[在线]。可用的是：https://arxiv.org/abs/1511.05641

[75] C.瑞士，W。刘，Y。贾，P。塞尔马内特，S.里德，D。安圭洛夫，D。厄尔汉，五世。范胡克和拉比诺维奇的《卷积更深入》。IEEE的续称。*计算机视觉与模式识别，2015年，第3页。* 1－9.

[76] R.Reed，“修剪算法——一个调查”，IEEE反式。*神经网络。*，第1卷。4、没有。第5页。740－747, 1993.

[77] J.冯和T。达雷尔，“学习深度卷积网络的结构”，在
该项目。内部信息。续称。美国计算机视觉出版社，2015年，第3页。2749－2757.

[78] J.杨，J。妈妈，贝里曼先生和P。佩雷斯，“一种大规模数据集的神经网络结构优化算法”。*2014年IEEE国际公司。续称。模糊的计算机系统。(FUZZ-IEEE)，2014年，第页。* 956－961.

[79] D.惠特利，T.。，“遗传算法和神经网络：优化连接和连接”，并行组合。，第1卷。14岁、没有。第3页。347－361, 1990.

[80] K.O.斯坦利和R。，“神经网络拓扑的有效进化”，在Proc中。*康格尔公司。进化计算(CEC02)，2002年，第2页*。1757－1762.

[81] T.真崎和S。渡边，“基于进化算法的深度神经网络的结构发现”，在Proc中。*2015年IEEE国际公司。续称。声学、语音和信号处理(ICASSP)，2015年，第页*。 4979－4983.

[82] B.沙赫里亚里，K。斯韦尔斯基，Z.。王先生，R。P.亚当斯，和N。德·弗雷塔斯，“让人类脱离循环：贝叶斯优化的回顾”，Proc。IEEE，第1卷。104年，没有。第1页。148－175，2016.

[83] D.拉马钱德拉姆，M.。莉丝琪，T。希尔兹，m.阿默尔，和G。泰勒，“使用图诱导核的深度多模态融合网络的结构优化”，在Proc中。*第25届欧洲音乐教学大纲。人工神经网络，计算智能和机器学习(ESANN)，布鲁日，比利时，2017，页*。 11－16.

[84] R.S.*萨顿和巴托，《强化学习》，卷。1.剑桥，麻省理工学院出版社，1998年*。

[85] B.Zoph和Q。V.Le。(2016). 神经结构搜索与强化学习。. ArXiv[在线]。可用的是：https://arxiv.org/abs/1611.01578

[86] P.库尔卡尔尼，J。西佩达，F。朱丽丝，P。佩雷斯，和L。*在Proc中，“使用L1正则化来学习深度架构的结构”。英国机器视觉股份有限公司。*，2015年，第3页。 23.1－23.11.

[87] N.斯里瓦斯塔瓦，G.欣顿，a.克里日耶夫斯基，我。“辍学：一种防止神经网络过拟合的简单方法，”J。*马赫数。学习经验。*，第1卷。15岁、没有。第1页。1929-1958年，1月1日。 2014.

[88] C. 默多克，李，周志熙，和T。*杜埃里格，“封锁：层次深度网络的动态模型选择”。内部信息。续称。计算机视觉与模式识别，2016年，第3页*。 2583－2591.

[89] F.李，N。无娃，C。狼，和G。泰勒，“模模：通过随机正则化学习多模态架构”，在Proc。*2017年IEEE续言。自动人脸和手势识别，2017年，第3页*。 422－429.

[90] M.第一，“RGBD数据集：过去、现在和未来”。*CVPR大规模三维数据研讨会：采集、建模和分析，2016年*。

[91] A.盖革，P。伦茨，C。“视觉与机器人技术相遇：Kitti数据集”，Int。*J.机器人学的经验。*，第1卷。32岁、没有。第11页。 1231－1237，2013.

[92] W.马德恩，G.帕斯科，C.里内加，和P。纽曼，“1年，1000公里：牛津机器人汽车数据集”，国际工业大学。*J.机器人学的经验。*，第1卷。36岁、没有。第1页。 3－15，2017.

[93] C.陈，R。贾法里，和N。凯塔纳瓦兹，“Utd-MHAD：利用深度相机和可穿戴惯性传感器进行人类动作识别的多模态数据集”，在Proc中。*2015年IEEE国际公司。Conf图像处理(ICIP)，2015年，第页*。 168－172.

[94] S.Escalera，X.巴罗，J。冈萨雷斯先生，鲍蒂斯塔先生，马达迪先生，先生。雷耶斯，V号。庞塞-洛佩兹，H。J.埃斯卡兰特等人，“2014年人们挑战结果：数据集和数据集”。*在欧洲委员会举办的研讨会。美国计算机视觉出版社，2014年，第3页*。459－473.

[95] F.奥弗里，乔德里，库里洛，维达尔和R。巴西，“伯克利MHAD：一个全面的多模式人类行动数据库”，在Proc。*2013年IEEE计算机视觉应用研讨会，2013年，第页*。 53－60.

[96] A.巴勃罗，Y州。莫拉德，F。戈莱莫的人。 C.穆里洛先生。洛普斯，和J。Civera，“一个多模态的人机交互数据集”，在Proc中。神经信息处理系统出版社，2016年，第3页。 1－5.

[97] F.林格瓦尔，A。桑德雷格，J。索尔，和D。拉兰，“介绍远程协作和情感互动的Recola多模态语料库”，在Proc中。*自动面部和手势识别第10个IEEEInt。续称。研讨会，2013年，第3页*。1－8.

[98] O.巴诺斯，C。维拉隆加，R。加西亚，A。萨伊斯，M。达马斯，J。 A.霍尔加多州。李安博，H。波马雷斯，和我。Rojas，“移动健康应用程序敏捷开发的新开放框架的设计、实现和验证，”生物医学工程师。*在线版，第1卷*。14岁、没有。2，p.S6,2015.[在线的]。可用的是：https://doi.org/10.1186/1475925X-14-S2-S6

[99] J.毛，J。徐，k、静、a、l。Yuille，“训练和评估与大规模网络注释图像的多模式单词嵌入”，在Proc。《神经信息处理系统的研究进展》，2016年，第3页。 442－450.

[100] J.阿雷瓦洛，T。索罗里奥先生。戈麦斯和F。A. 冈萨雷斯。(2017). 用于信息融合的门控多模态单元。. ArXiv[在线]。可用的：https://arxiv.org/abs/1702.01992

[101] Y.-G.江，Z。吴，王俊强，X。薛和南f。Chang，“利用正则化的深度神经网络开发视频分类中的特征和类关系”，IEEE反式。*模式的肛门。马赫数。被告知。*，2017. 多：https://doi.org/10.1109/TPAMI.2017.2670560

[102] R.Min，N.Kose和J.-L.Dugelay，“KinectFaceDB：人脸识别的Kinect数据库”，IEEE反式。*系统系统。，人，西伯恩。，系统。*，第1卷。44岁、没有。第11页。1534-1548年，11月。 2014.

[103] B.H.Menze，a.s。鲍尔，J。卡尔帕蒂-克莱默，法拉哈尼。科比，Y州。伯伦，N。Porz等人，“多模态脑肿瘤图像分割基准(brats)”，IEEE反式。*地中海的人。伊玛目集团。*，第1卷。34岁、没有。第10页。 1993－2024，2015.

sp的值