第 1 章

绪　论

1.1　概述

近年来，深度学习算法在视觉处理、语音识别及自然语言处理等领域取得了重大突破，但当同时面对多种来源的多模态数据时，其处理能力已接近瓶颈。事实上，人工智能领域的众多应用场景往往需要输入多种模态数据，譬如个人智能助理系统需要全面理解融合了口语、肢体语言及视觉信息的混合输入。因此，深入探索跨越多种模态的建模与训练方法，实现多模态数据的表征、融合、理解与推理，对持续推动多模态智能信息处理技术的革新与发展，具有划时代意义。

定义 1.1（多模态智能信息处理） 多模态智能信息处理巧妙地将来自文本、图像、语音、视频等多种传感器和数据源的丰富信息整合在一起，进行深入的分析和处理。这项技术融合了计算机视觉、自然语言处理、音频处理，以及机器学习等多种技术，能够从不同的感知渠道中获取更加全面、细致的信息。利用这些来自多种感知模态的信息，多模态智能信息处理技术为各个领域的智能系统提供了更加广泛、深入的数据理解模式，以及更加强大、灵活的信息交互能力。这不仅让智能系统变得更加聪明、更加懂你，也为我们的生活和工作带来了更多的便利和惊喜。通过多模态表示学习，我们能够更好地理解和表示这些多样化的信息；多模态信息融合则让这些信息相互补充、协同工作；多模态信息检索让我们能更快速地找到所需的信息；多模态内容理解让机器能更深入地“读懂”这些信息的内涵；而多模态视觉生成更是让机器能够创造丰富多彩的视觉内容。 ♣

1.2　理论基础

1.2.1　卷积神经网络

卷积神经网络（Convolutional Neural Networks，CNN）是一种广泛应用于深度

学习的通用模型。它具备处理多维度数据的能力，无论是音频、图像还是视频，都能应对自如。该模型通过巧妙地堆叠卷积层、池化层及全连接层，逐层提取数据特征，以完成分类、识别、检测等多重任务。凭借其卓越的性能，卷积神经网络广泛应用于计算机视觉、自然语言处理及医学影像分析等诸多领域。随着持续的创新与技术改进，卷积神经网络已成为人工智能领域最具影响力的模型之一。

在这里，我们以利用双模态卷积神经网络处理 RGB-D 数据[①]对为例，进行详细介绍。如图1.1 所示，RGB 图像和其对应的深度信息被分别输入两个独立的卷积神经网络中，经过处理后输出各自的特征表示。对于当前的空间位置 i，其融合后特征 $\boldsymbol{f}_i^{\text{fusion}}$ 由单模态特征加权求和得到：

$$\boldsymbol{f}_i^{\text{fusion}} = \sum_{j=1}^{N} \boldsymbol{w}_i^j \boldsymbol{f}_i^j \tag{1.1}$$

其中，$\boldsymbol{f}_i^j$ 是第 j 个模态的特征图，$\boldsymbol{w}_i^j$ 表示权重向量，可由下式计算得到：

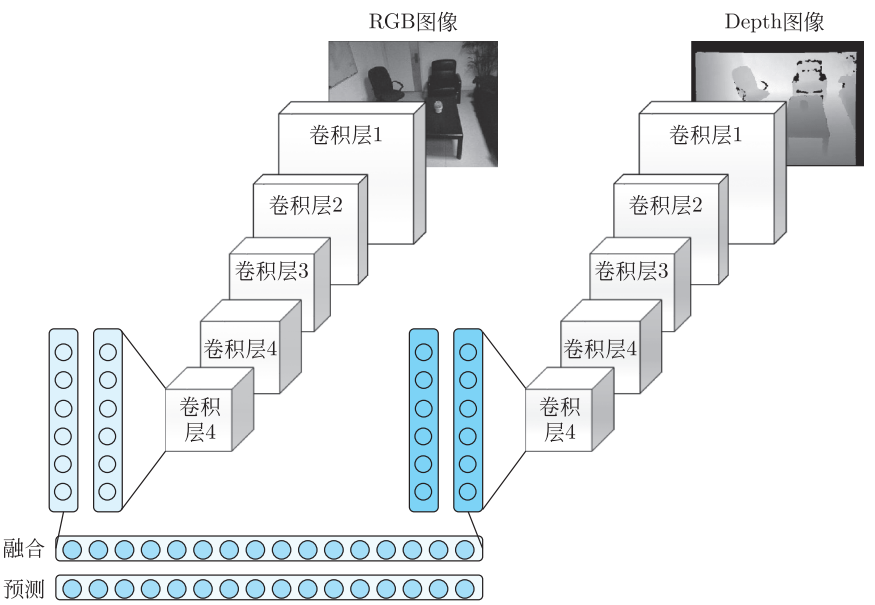


图 1.1　双模态卷积神经网络示例

① RGB-D 数据是利用传感器采集到的包含彩色图像及其深度信息的具有四维通道的数据。

$$\boldsymbol{w}_i^j = \frac{\exp\left(\boldsymbol{f}_i^j\right)}{\sum_{k=1}^{N} \exp\left(\boldsymbol{f}_i^k\right)} \tag{1.2}$$

作为一种强大的特征提取器，多模态卷积神经网络可以从不同模态中学习局部跨模态特征。此外，它还能够从多模态数据流中对空间信息进行建模，大大扩展了其应用领域。如图1.2 所示，在多模态数据驱动的目标分割任务中，半监督多模态分割方法（Semi-CML）能够利用少量标记数据实现准确的分割。该框架利用大量未标记的多模态数据来提高每种模态的分割性能，并确保在推断阶段只使用一种模态就可进行准确的医学图像分割。

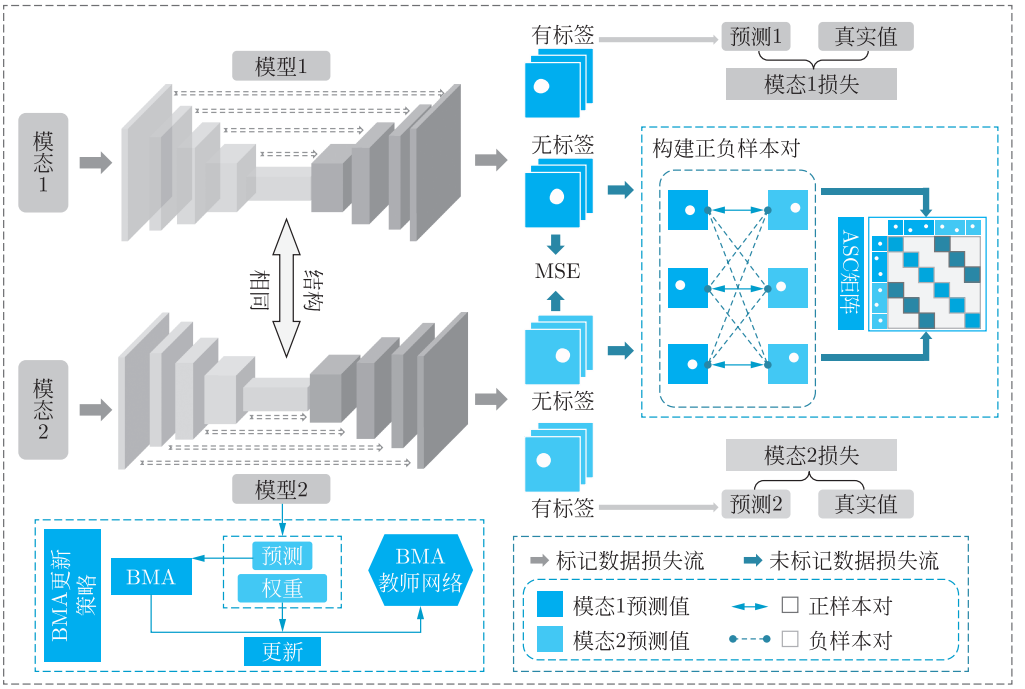


图 1.2　半监督多模态分割网络框架

小　贴　士

卷积神经网络是多层感知机的变种，由生物学家 David Hubel 和 Torsten Wiesel 在早期关于猫视觉皮层的研究发展而来。该研究发现视觉皮层的细胞

存在一个复杂的构造，使得这些细胞对视觉输入空间的子区域非常敏感，这一子区域被称为感受野。初代卷积神经网络是由美国纽约大学的 Yann Lecun 教授于 1998 年提出的 LeNet-5，本质上仍是一个多层感知机。其所采用的局部连接和权值共享的方式减少了权值的数量，使得网络易于优化，同时也降低了模型的复杂度、减少了过拟合的风险。当网络的输入为图像时，上述优点体现得更为明显。2006 年，Geoffrey Hinton 首次提出“深度学习”这一概念，其主要观点是：多隐层的人工神经网络具有优异的特征学习能力，学习到的数据更能反映数据的本质特征，有利于可视化或分类。随着大数据和计算机硬件的迅速发展，深度学习得以大范围推广和应用。2012 年，由 Geoffrey Hinton 等提出的 AlexNet 取得 ILSVRC-ImageNet 竞赛分类任务的冠军，至此，卷积神经网络的研究与应用迎来爆发式增长。鉴于 John Hopfield 与 Geoffrey Hinton 开创性地利用物理学工具来训练人工神经网络，促成了当今深度神经网络及深度学习技术的诞生，2024 年这两位科学家被授予诺贝尔物理学奖。

1.2.2 池化

池化（Pooling）作为一种常用的神经网络层可以用来实现多模态信息融合。池化层通常紧随卷积层，可以减少特征图的维度和参数数量，从而提高模型的计算效率和鲁棒性，有助于保留图像或序列数据中的关键信息，减小过拟合风险，并使模型对平移和旋转等变换具有更好的不变性，在卷积神经网络中起到至关重要的作用。

池化的发展历史可以追溯到 20 世纪 80 年代末和 90 年代初。最早的池化方法则可以追溯到卷积神经网络的先驱模型 LeNet。该模型采用平均池化的方式来减小特征图的尺寸，并保留主要信息。随着深度学习技术的兴起，一系列改进的池化方法相继提出以适应不同任务和数据类型的需求。目前主流的池化方法有以下几种：（1）平均池化（Average Pooling），是最早被广泛采用的一种池化方法，如图1.3所示，它通过计算每个池化窗口内的特征值的平均值来减少特征图的维度；（2）最大池化（Max Pooling），是一种常见的池化方法，通过选择池化窗口内的最大值来生成新的特征图，有助于保留图像中的显著特征；（3）自适应池化（Adaptive Pooling），是一种根据输入尺寸动态调整输出尺寸的池化方法，可根据输入的形状自动调整输出大小，避免了固定大小池化的局限性，提高了模型的灵活性和泛化能力；（4）分层池化

（Hierarchical Pooling），在处理多尺度信息或分层结构的数据时，引入分层池化，以减小特征图的维度。分层池化将不同层次的特征按照层级进行池化，使模型能够更好地处理多尺度信息，提高特征学习的鲁棒性。

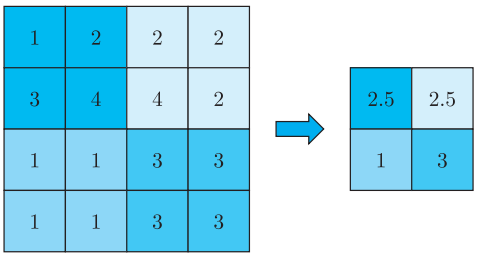


图 1.3　平均池化

1.2.3　注意力机制

小　贴　士

“Attention is All Your Need”是由谷歌机器翻译团队于 2017 年发表于 *NeurIPS* 的一篇文章。该论文最大的贡献在于抛弃了以往机器翻译普遍采用的循环神经网络或卷积神经网络等传统架构，以编码器-解码器为基础，创新性地提出了 Transformer 架构。该架构可以有效地解决卷积神经网络无法并行处理及卷积神经网络无法高效捕捉长距离依赖的问题，近期更是被进一步地应用到了计算机视觉领域，同时在多个计算机视觉任务上取得了最优性能，挑战卷积神经网络在计算机视觉领域多年的霸主地位。

注意力机制于 2014 年被首次提出，当时 Yoshua Bengio 等将注意力机制引入神经网络模型，并应用于机器翻译任务。他们提出的基于内容的注意力机制，可以根据输入序列的不同部分动态地调整输出序列的生成过程，从而提高机器翻译的准确性和流畅度。这一模型被称为“注意力神经机器翻译”（Attention-based Neural Machine Translation），开创了注意力机制在深度学习中的应用先河。随着时间的推移，多种不同类型的注意力机制嵌入各种深度学习模型中。早期的注意力机制主要应用于“序列到序列”模型中，解决文本翻译等任务。在这些模型中，注意力机制主要用于选择源序列中与目标序列相关的部分，以辅助输出序列的生成。通过这种方式，模型可以更好地处理长序列数据和建立源目标之间的关联。

注意力机制可以让模型关注特征图的特定区域或特征序列的时间步长。通过注意力机制，不仅可以提升模型性能，还能增强特征表示的可解释性。该机制模拟了人类提取最具辨别力信息的能力。注意决策过程不是一次使用所有信息，而是有选择性地聚焦于场景中与任务最相关的部分。这一机制在视觉分类、神经机器翻译、语音识别、图像描述等许多应用中展现了强大能力。

根据在选择部分特征时是否使用键（Key），注意力机制可以分为两类：基于键的注意力和无键注意力。在基于键的注意力机制中，使用键来搜索显著的局部特征。以图像描述任务为例，其典型结构如图1.4所示，其中卷积神经网络将图像编码为特征集 $\boldsymbol{a}_i$，然后循环神经网络将输入解码为隐藏状态 $\boldsymbol{h}_t$。在时间步 t，输出 $\boldsymbol{y}_t$ 是基于 $\boldsymbol{a}_t$ 和 $\boldsymbol{c}_t$ 计算得到的，其中，$\boldsymbol{c}_t$ 是提取的显著特征。在提取 $\boldsymbol{c}_t$ 的过程中，解码器当前的状态 $\boldsymbol{h}_t$ 起着重要作用，因为它决定了应该关注哪些特征，而编码器的状态 $\boldsymbol{a}_i$ 则是提供这些特征的来源。简单来说，解码器利用当前状态 $\boldsymbol{h}_t$ 去“挑选”编码器中的特征 $\boldsymbol{a}_i$，以便更好地预测输出 $\boldsymbol{y}_t$。在无键注意力机制中，则不使用额外的键向量来引导特征选择，而是直接基于特征本身进行聚合。例如，可以对所有特征 $\boldsymbol{a}_i$ 赋予相同的权重，或使用固定权重策略来提取上下文向量 $\boldsymbol{c}_t$。这类机制更为简单，但缺乏对当前解码状态 $\boldsymbol{h}_t$ 的动态适应能力，因而在建模复杂依赖关系时可能表现不如基于键的注意力。

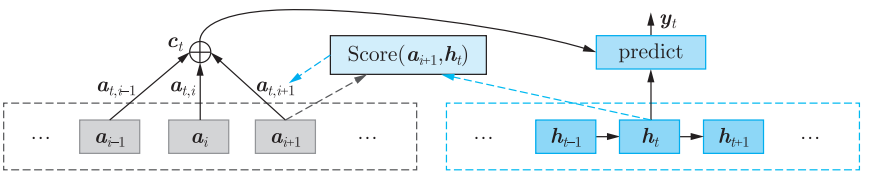


图 1.4　基于键的注意力机制典型结构

基于键的注意力在视觉描述应用中很普遍，通常采用编码器-解码器网络，这为我们带来了一种评估模态内或模态之间特征重要性的方法。一方面，注意力机制可用于选择模态内最显著的特征；另一方面，它可用于在融合多个模态时平衡模态之间的贡献度。

为了识别和描述视觉模态中包含的对象，一组对不同对象进行不同编码的局部区域特征比单个特征向量更有帮助。通过动态选择图像中最显著的区域或视频序列的时间步长，可以提高系统性能和噪声鲁棒性。注意力机制可以用于自动检测图像中的重要区域（如图1.5所示），并将这些视觉特征与解码器中的文本特征相结合，以

得到图像的字幕。在每个时间步 t 上，注意力模块会根据当前生成的文本自动搜索适合预测下一个单词的局部区域。

（a）一个女人在公园扔飞盘　（b）一只狗趴在木地板上　（c）一条路上的停车标识，背景是一座山

（d）一个小女孩和她的泰迪熊坐在床上　（e）一群人坐在船上　（f）一只长颈鹿站在森林里，背景是树

图 1.5　注意力机制关注到正确目标的示例

（注：每幅图右侧光圈部分表示关注区域，下画线表示相应的目标）

为了更精确地定位图像中的局部特征，研究人员提出了多种注意力模型并持续推动其发展。堆叠式注意力网络通过多步搜索的方式，在每个时间步使用语言特征作为键值，引导注意力在图像中寻找显著区域。每次搜索后，模型将注意力集中的视觉和语言特征组合成一个向量，并将其作为下一步迭代的输入。随着迭代的深入，模型不仅能够在多个步骤中逐步定位更加细粒度的局部区域，还可以有效融合视觉和语言特征。此外，结构化注意力模型可以捕捉图像区域之间的语义关系，进一步提升模型对复杂图像结构的理解能力。该模型通过分析图像区域间的空间布局来增强注意力的选择性，使系统能够更好地推断各区域的相对位置与联系，进而在复杂场景中准确地将注意力置于正确区域。此外，空间和通道注意力机制也被整合到卷积神经网络中，通过过滤和强化不同尺度的特征信息，提高了模型在复杂视觉任务中的性能。如图 1.6 所示，空间和通道注意力机制让模型不仅能够关注局部区域特征，还能够处理卷积网络中的不同通道特征。

近年来，作为注意力机制的一个重要扩展，多头注意力（Multi-Head Attention）机制得到广泛关注与应用。它利用多个注意力模块从相同的输入数据中提取不同类型的特征，增强模型的表达能力和学习效率，从而更全面地捕获特征和上下文信息。通常情况下每种类型的特征位于不同的子空间中，并代表不同的语义。因此，使用多头注意力机制可以发现模态之间的不同交互情况。通过这种方式，多头注意力机制能够有效地整合多模态信息，从而提升模型的表达能力。

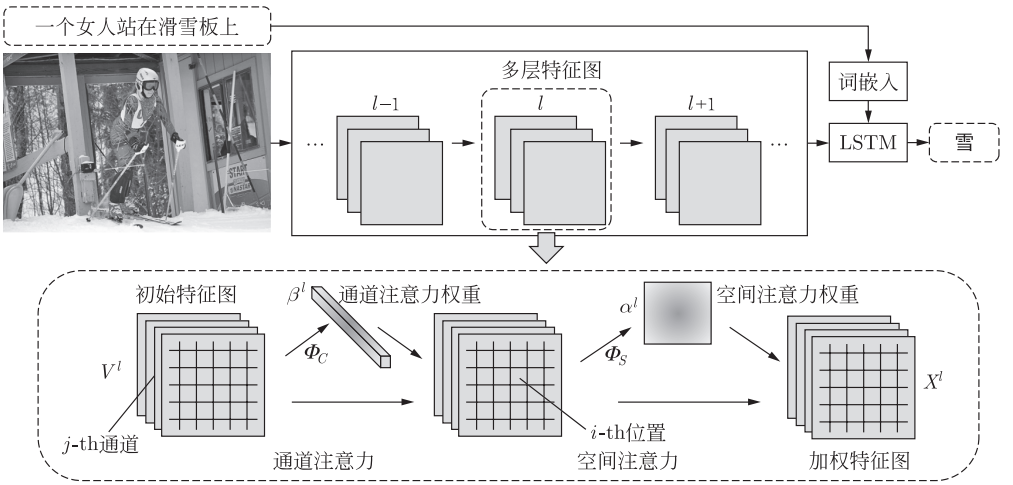


图 1.6　空间和通道注意力机制

1.2.4　Transformer 网络

小　贴　士

Transformer 网络于 2017 年横空出世，当时被定位成一种简单并且可扩展的自然语言翻译方法，并且很快地应用到各类 NLP 任务中，逐渐成为高性能模型中的必备组件。同时，Transformer 网络也为视觉领域带来了革命性的变化，它让视觉领域中的目标检测、视频分类、图像分类和图像生成等多个任务有了长足的进步。高效结合 Transformer 网络的新模型性能可以轻松超越该领域现有的最优模型。

Transformer 网络引入自注意力机制，使得模型能够并行地处理序列数据，并且具有较强的建模能力和泛化能力。通常，Transformer 网络将 RGB 图像（或音频频谱图）分割成 N 个不重叠的块（patch），然后利用线性投影将它们转换为一维词元（token）。同时还会额外引入一个向量，作为分类任务的特征，以及一个可学习的位置向量，用来表示输入特征的位置信息。然后将得到的词元输入由 L 个 Transformer 层组成的编码器中。每个 Transformer 层由多头自注意力（Multi-head Self-Attention，MSA），层归一化（Layer Normalization，LN）和多层感知机（Multi-Layer Perception，MLP）等部分组成。图1.7给出了一种经典的 Transformer 网络框架。

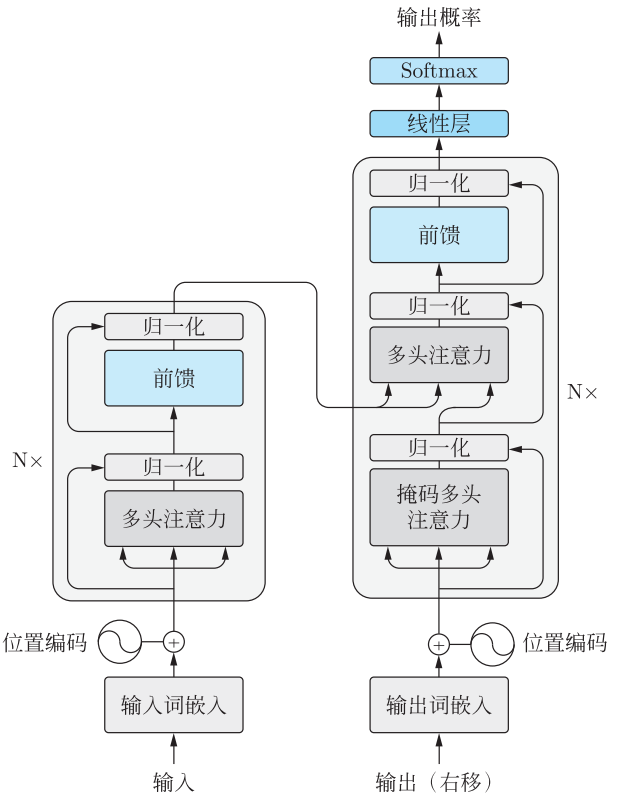


图 1.7　Transformer 网络框架

1.3　技术难点与挑战

多模态智能信息处理涉及从不同数据模态（如视觉、文本、语音等）中提取、融合和生成信息的复杂过程。其面临的主要技术挑战可以概括为三方面。

（1）**输入处理**：如何有效获取不同类型的输入数据并进行编码，使得各模态数据能够被合理地表示和处理。由于每种模态具有其独特性，传统的手工特征提取方式难以满足多模态任务的需求，而深度学习提供了自动学习特征的能力，能更好地应对异构数据输入的复杂性。

（2）**信息融合**：如何将来自不同模态的特征融合到统一的表示空间中。早期的简单融合方法存在信息损失和交互不足等问题，而更先进的融合方法，如基于自注意力机制的 Transformer，能够更有效地对齐和融合多模态信息。

（3）**输出生成**：如何根据融合后的信息进行预测、决策或生成内容。由于多模态

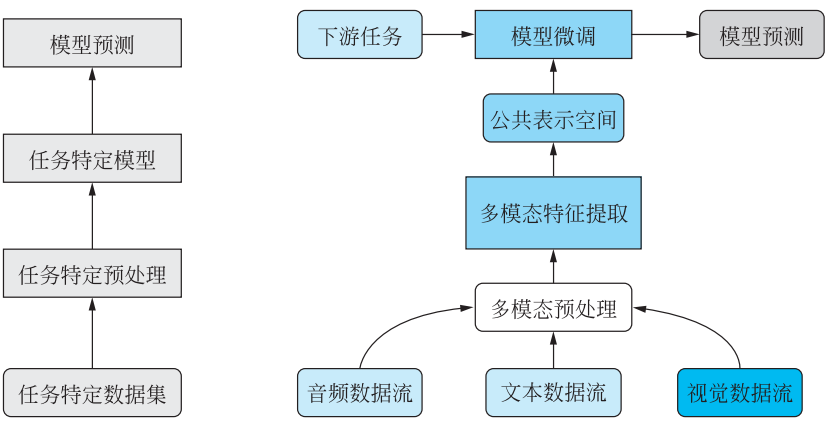
信息处理系统通常涉及复杂的决策过程，模型需要具备处理多模态数据的泛化能力和灵活性。

多模态智能信息处理系统的技术难点主要集中在上述三个方面，下文主要介绍前两个方面。除此之外，可解释性、安全性等方面也面临诸多挑战。

1.3.1 输入处理

在处理多模态输入时面临的第一个挑战是理解不同模态的编码，即学习表示中的异质性差异。在实际应用中，我们可以通过关联不同模态的相似语义来减少此差异，并融合来自不同模态的特征以提高跨模态任务的性能。得益于强大的学习能力，深度学习在理解和处理表征方面优于经典的浅层学习方法，特别是在没有附加信息的情况下，可以将不同模态的信息转化为通用特征。因此，深度学习可以很好地兼容多模态表示学习，处理多模态输入。

适用于特定任务的多模态方法遵循流水线模式，通常仅能处理一个任务，如图1.8（a）所示。例如，专为视觉-语言（Vision-Language，VL）任务设计的方法基于卷积神经网络的视觉编码器提取视觉特征，还有基于词袋、序列模型或转换器的文本编码器来编码文本特征。随后，多模态融合在不同模态的特征之间建立交互以产生通用表示。这些多模态融合方法演变为如下几种范式。



（a） 基于特定任务的多模态方法　　（b） 基于多模态交互融合的多模态方法

图 1.8　特定于任务的多模态方法

（1）简单融合：通过级联或逐元素求和或乘积融合特征。

（2）模态间注意力机制：捕获不同模态的特征对齐并构建更多信息联合表示。

（3）模态内注意力机制：通过构建图形表示来发展关系推理。