

# 跨模态人工智能：融合视觉与语言的新趋势

岳泽宇

微软（中国）有限公司 北京 100080

**摘要：**随着人工智能技术的飞速发展，单一模态的智能系统已难以满足复杂多变的应用需求。跨模态人工智能，尤其是视觉与语言的融合，成为当前研究的热点与前沿方向。本文旨在探讨跨模态人工智能中视觉与语言融合的新趋势，分析其理论基础、关键技术、应用场景及未来发展方向，以期对相关领域的研究与实践提供参考。

**关键词：**人工智能；多模态；视觉；语言

## 引言

在人工智能领域，模态指的是智能体接收和输出信息的特定方式，主要包括语音、文本、图像、视频等。近年来，基于深度学习技术的单一模态研究，如计算机视觉、自然语言处理等，取得了显著进展。然而，面对更高层次的人工智能任务，单一模态的局限性日益凸显，跨模态信息处理成为必然趋势。视觉与语言的融合作为跨模态研究的重要组成部分，具有广阔的应用前景和研究价值。

## 1 跨模态人工智能的理论基础

### 1.1 多模态学习的概念

多模态学习旨在利用不同感官通道获取的多种形式的信息，如视觉、听觉、触觉等，通过跨模态的特征提取、关联建模和联合优化，实现比单一模态更好的学习效果。其核心思想是，通过学习和融合不同模态数据之间的内在联系，获得更加丰富和准确的语义表示，从而提升智能系统的整体性能。多模态学习不仅仅是对多个模态数据的简单叠加，更是对这些数据进行深度整合和挖掘，以实现不同模态之间的互补和增强<sup>[1]</sup>。例如，在语音识别任务中，结合音频信号和视觉信息（如说话人的面部表情和手势）可以显著提高识别的准确率。

### 1.2 跨模态表征学习

跨模态表征学习是将多个模态数据所蕴含的语义信息投影到连续向量表征空间，以实现信息融合和推理。这一领域的研究重点在于如何构建跨模态的统一语义空间，将不同模态的特征进行有效整合。这要求模型能够捕捉到不同模态数据之间的共有语义特征，同时保留各自模态的独特信息。为了实现这一目标，研究者们提出了多种方法，包括共享表示学习、协同表示学习和对抗性学习等。近期的研究如VL-BERT等模型，在图像/视频与文本融合方面取得了显著进展，为跨模态表征学习提供了新的思路和方法。VL-BERT通过引入Transformer架

构，实现了图像和文本数据的联合建模，使得模型能够同时处理和理解这两种模态的信息。此外，该模型还采用了预训练的策略，通过大量图像-文本对的学习，提升了模型在跨模态任务上的泛化能力。这些研究成果不仅推动了跨模态表征学习的发展，也为跨模态人工智能的应用提供了有力的支持。

## 2 跨模态人工智能的关键技术

### 2.1 特征提取与表示

不同模态数据具有各自的物理特征和语义特征，如何有效地提取和表示这些特征是多模态学习的基础。在跨模态人工智能中，特征提取与表示技术起着至关重要的作用，它要求模型能够从不同模态的数据中提取出有用的信息，并将其转化为一种统一的、可比较的格式。对于图像数据，卷积神经网络（CNN）是一种常用的特征提取方法。CNN通过卷积层、池化层和全连接层等结构，能够自动学习到图像中的边缘、纹理、形状等低级特征，以及更高级别的语义特征，如物体、场景等。这些特征对于后续的跨模态关联建模和任务优化具有重要意义。对于文本数据，循环神经网络（RNN）和Transformer是两种主流的特征提取方法。RNN通过捕捉序列信息中的时间依赖性，能够有效地处理自然语言文本。它能够将文本转化为一组向量表示，每个向量都包含了该位置单词及其上下文的信息。而Transformer则采用了一种自注意力机制，能够更好地处理长距离依赖关系，并在多个任务上取得了显著的效果。除了CNN、RNN和Transformer，还有其他一些技术也常用于多模态特征提取，如用于音频信号处理的深度神经网络（DNN），以及用于处理时间序列数据的长短时记忆网络（LSTM）等。这些技术能够根据不同模态数据的特性，提取出相应的特征表示，为后续的多模态融合和推理提供基础。

### 2.2 跨模态关联建模

跨模态关联建模是多模态学习的核心环节,旨在建立不同模态数据之间的内在联系,以实现信息的有效融合和推理。常用的方法包括协同注意力机制、跨模态对比学习、多模态融合网络等,这些方法在不同的应用场景下展现出了强大的建模能力。协同注意力机制是一种有效的跨模态关联建模方法,它通过计算不同模态数据之间的注意力权重,实现模态间的信息交互和增强。具体来说,协同注意力机制可以利用一种模态的数据作为查询,去检索另一种模态中与之相关的信息,从而实现跨模态的关联和对齐。跨模态对比学习则是一种通过比较不同模态数据之间的相似性和差异性来建立关联的方法。它通常将不同模态的数据映射到一个共同的语义空间,并在这个空间中计算它们之间的距离或相似度。通过优化这个距离或相似度,模型可以学习到不同模态数据之间的内在联系和对应关系<sup>[2]</sup>。多模态融合网络则是将不同模态的数据进行融合,并利用深度学习模型进行联合建模和优化的方法。这种网络通常包含多个子网络,每个子网络负责处理一种模态的数据,并通过融合层将不同模态的特征进行融合。通过端到端的训练,多模态融合网络可以学习到不同模态数据之间的复杂关联和交互。这些方法在跨模态人工智能中发挥着重要作用,它们通过建立复杂的跨模态关联,增强了系统对事物的理解能力,提高了模型的鲁棒性和准确性。在实际应用中,这些方法可以根据具体的任务和数据特性进行选择 and 组合,以实现最佳的跨模态学习效果。

### 2.3 联合优化与训练策略

在多模态学习中,联合优化是一个至关重要的环节,它要求在特征提取、关联建模、任务优化等多个环节进行端到端的整体优化,以达到最佳的学习效果。为了实现这一目标,需要设计复杂的神经网络架构,并采用一系列合适的训练策略。在神经网络架构方面,研究者们通常采用深度神经网络(DNN)作为基本框架,并结合具体的多模态任务进行定制化的设计。这些网络架构往往包含多个层次,用于逐步提取和融合不同模态的特征。例如,在图像和文本的多模态任务中,可以使用卷积神经网络(CNN)来提取图像特征,使用循环神经网络(RNN)或Transformer来提取文本特征,并通过融合层将这些特征进行有效整合。在训练策略方面,多任务学习是一种常用的方法。它通过同时优化多个相关任务来提高模型的泛化能力。在多模态学习中,可以将不同的模态任务视为相关的子任务,并通过共享表示层或联合损失函数来进行多任务学习。这样,模型可以在不同任务之间共享有用的信息,从而提高整体性能。此

外,迁移学习也是多模态学习中常用的训练策略。它利用已有的预训练模型作为起点,通过微调(fine-tuning)或特征提取的方式,将知识迁移到新的多模态任务上。迁移学习可以有效利用大量未标注的多模态数据进行预训练,提高模型的泛化能力和学习效果。

## 3 跨模态人工智能的应用场景

### 3.1 视觉问答(VQA)

视觉问答(Visual Question Answering, VQA)是一种结合了计算机视觉和自然语言处理技术的跨模态任务。该任务要求系统根据输入的图像内容回答用户提出的自然语言问题。视觉问答系统不仅需要准确理解图像中的视觉信息,如物体、场景、动作等,还需要理解问题的语义,并将两者有机结合以生成准确的答案。应用场景包括:(1)早期教育:在教育领域,视觉问答系统可以作为辅助工具,帮助低年级学生更好地理解图像内容,提高他们的观察力和语言表达能力。(2)辅助视觉障碍人士:对于视觉障碍人士,视觉问答系统可以帮助他们通过语音指令获取图像信息,了解周围环境,提高生活自理能力。(3)智能客服:在电商、医疗等领域,智能客服系统可以结合视觉问答技术,通过用户上传的图片和问题,提供更加直观和准确的解答。

### 3.2 图像/视频转文字(Captioning)

图像/视频转文字任务,也称为图像/视频描述(Image/Video Captioning),要求系统能够自动生成描述图像或视频内容的文本。这一任务不仅考验系统对视觉信息的理解能力,还要求其具备将视觉信息转化为自然语言的能力。应用场景包括:(1)社交媒体:在社交媒体平台上,用户上传的图片和视频可以自动生成描述性文本,提高内容的可读性和分享性。(2)新闻报道:新闻机构可以利用图像/视频转文字技术,自动生成新闻图片或视频的描述,加快新闻报道的速度和效率。(3)视频制作:在视频制作领域,该技术可以用于自动生成视频字幕,提高视频的可访问性和国际化水平。

### 3.3 跨模态检索

跨模态检索允许用户通过一种模态的信息(如文本)检索到另一种模态的信息(如图像、视频或音频)。这种检索方式打破了传统单一模态检索的局限性,实现了更加灵活和高效的信息获取方式。应用场景包括:(1)基于文本描述的图像检索:用户可以通过输入描述性文本,检索到与之相关的图像资源。这在电商、旅游、医疗等领域具有广泛应用,如通过描述搜索商品图片、旅游景点图片或医学图像等<sup>[3]</sup>。(2)基于视频内容的语音识别:在视频处理领域,跨模态检索技术

可以实现视频内容的语音识别和关键词检索,方便用户快速定位视频中的关键信息。(3)多媒体事件检测:在网络舆情监测、公共安全等领域,跨模态检索技术可以用于检测和分析多媒体事件,提高事件的响应速度和处理效率。

#### 4 跨模态人工智能的未来发展方向

##### 4.1 跨模态统一建模

随着人工智能技术的不断发展,单一模态的处理能力已难以满足复杂任务的需求。跨模态统一建模旨在构建一个能够处理多种数据模态(如文本、图像、视频、音频等)的通用模型,以应对现实世界中多模态信息的复杂性和多样性。为了实现跨模态统一建模,需要在模型架构上进行创新。例如,采用Transformer等先进的神经网络架构,结合多模态预训练技术,构建一个能够同时处理多种模态数据的通用模型。这种模型将具备更强的跨模态语义对齐能力,能够更有效地融合不同模态的信息。跨模态统一建模的关键在于实现跨模态语义对齐。这要求模型能够理解不同模态数据之间的内在联系和语义关系,实现模态间的无缝转换和融合。为了实现这一目标,需要研究跨模态表征学习、跨模态信息融合等关键技术。跨模态统一建模需要大规模、高质量的多模态数据集进行训练。同时,引入领域知识、常识知识等外部信息,可以进一步提升模型的跨模态理解和生成能力。

##### 4.2 高级认知智能

随着人工智能技术的普及和应用,人们对其智能化水平的要求也越来越高。高级认知智能旨在使机器具备更接近于人类的思维方式,实现更加智能化、人性化的交互体验。常识知识是人类智能的重要组成部分,也是机器实现高级认知智能的关键。跨模态常识知识学习要求机器能够结合多种模态的信息进行常识推理和判断,更准确地理解现实世界中的复杂场景和任务。情感智能是人类特有的认知能力之一,也是人机交互中的重要环节。跨模态情感智能要求机器能够理解和模拟人类的情感表达,实现更加智能化、人性化的交互体验<sup>[4]</sup>。例如,在智能语音助手领域,跨模态情感智能可以使语音助手

更好地理解用户的情感状态和需求,提供更加贴心、个性化的服务。

##### 4.3 多场景下的多模态交互

随着人工智能技术的普及和应用场景的多样化,多模态交互已成为提升应用性能的关键。多场景下的多模态交互要求智能系统能够在各种复杂场景中灵活地运用多种模态的信息进行交互,以应对不同的任务和挑战。多场景下的多模态交互要求智能系统具备更强的复杂场景适应性。这要求系统能够根据不同的场景和任务需求,灵活地选择和组合不同的模态信息进行交互,以实现最佳的交互效果。多模态交互旨在提供更加自然、高效的人机交互体验。通过融合多种模态的信息(如语音、手势、表情等),智能系统可以更准确地理解用户的需求和意图,提供更加精准、个性化的服务。实现多场景下的多模态交互需要克服一系列技术挑战,如跨模态数据的同步与对齐、多模态信息的融合与推理等。未来,随着技术的不断进步和创新(如跨模态学习算法的优化、多模态融合技术的改进等),这些挑战将逐渐得到解决,推动跨模态人工智能向更高水平发展。

#### 结语

跨模态人工智能,尤其是视觉与语言的融合,是当前人工智能领域的重要研究方向。通过理论创新、技术突破和应用拓展,跨模态人工智能将为实现更高层次的人工智能任务提供有力支撑,推动人工智能技术的持续进步和发展。

#### 参考文献

- [1]廖俊淇,魏昕,周亮.人工智能驱动的跨模态语义通信系统[J/OL].中兴通讯技术,1-12[2024-07-25].
- [2]唐昆,李白杨,张心源.基于主客观融合的人工智能跨模态生成内容质量及效能测度研究[J/OL].情报理论与实践,1-15[2024-07-25].
- [3]吴岸雄,赵嘉凌,黄少伟,等.多模态人工智能数据分析实验服务平台建设[J].实验室研究与探索,2023,42(04):188-193.
- [4]多模态人工智能正大步走向场景应用新阶段[J].机床与液压,2022,50(19):147.