

多模态人工智能系统中的信息融合瓶颈与应对策略

李思睿

东北农业大学 黑龙江哈尔滨 150030

摘要：多模态人工智能系统旨在整合文本、图像、音频等多种不同类型的数据以提升智能系统的性能和表现，不过在信息融合过程中却面临着诸多瓶颈，像数据的异构性、数据对齐困难、计算资源需求大等问题，本文则详细分析了这些瓶颈并提出了相应的应对策略，目的是推动多模态人工智能技术的进一步发展。

关键词：多模态人工智能；信息融合；瓶颈；应对策略；统一特征空间

引言

随着人工智能技术的快速发展，多模态人工智能系统逐渐成为研究热点，其中多模态数据包含来自不同感知渠道的信息，例如视觉方面的图像、视频，听觉方面的音频，语言方面的文本等，通过融合这些多模态数据，人工智能系统能够更全面、准确地理解和处理信息，从而在自动驾驶、医疗诊断、智能客服等众多领域展现出巨大的应用潜力。

一、多模态人工智能系统中的信息融合瓶颈

（一）数据异构性问题

不同模态的数据存在天然的异构性，文本数据以离散的符号序列形式存在，其语义理解依赖于词汇和语法结构；图像数据则是连续的像素矩阵，包含丰富的视觉特征如颜色、纹理、形状等；音频数据是随时间变化的波形信号，反映声音的频率、振幅等特征，这些数据在表示形式、数据结构和特征维度上差异巨大，比如文本的词向量维度通常在几百维，而图像经过卷积神经网络提取的特征向量维度可能高达数千维，这种异构性使得直接对不同模态数据进行融合变得非常困难，模型难以从中学习到有效的关联信息，导致融合效果不佳。

（二）数据对齐难题

在多模态数据中，确保不同模态信息在时间或语义上的对齐至关重要，可实际应用中数据对齐面临诸多挑战，比如在视频分析里，音频和视频流需要在时间轴上精确同步，微小的时间偏差都可能导致信息理解的偏差，像电影字幕与音频的匹配中，字幕出现延迟或提前就会严重影响观众的观看体验^[1]；在图文理解任务中，文本描述与图像内容需要在语义层面准确对应，可由于文本

的模糊性和图像内容的复杂性，建立准确的语义对齐关系并不容易，例如对于一张包含多个物体的复杂图像，文本描述可能只提及部分关键物体，如何确定图像中其他物体与文本的关联便是实现语义对齐的难点；此外，不同模态数据的采集频率和更新周期往往不同，这进一步增加了数据对齐的难度。

（三）计算资源需求巨大

训练多模态人工智能系统需要处理海量多模态数据，且不同模态数据的处理通常需要不同类型的网络架构和计算流程，比如处理图像数据一般采用卷积神经网络（CNN），它通过卷积、池化等操作提取图像的局部特征，而处理文本数据常使用基于Transformer的架构以捕捉文本中的长距离依赖关系；同时运行这些不同类型的网络对计算资源的需求极为庞大，此外，为了实现多模态数据的有效融合，模型往往需要更深的网络结构和更多的参数，这进一步加剧了计算负担，在实际应用中，尤其是在资源受限的边缘设备上，满足如此巨大的计算资源需求成为阻碍多模态人工智能系统部署和应用的关键因素。

（四）模型训练的稳定性与可解释性差

多模态模型由于涉及多种模态数据的交互和复杂的网络结构，其训练过程的稳定性较差，在训练过程中容易出现过拟合现象，尤其是当某一模态的数据量相对较少或特征较为复杂时，模型可能过度学习这一模态的特定特征，而忽略了其他模态的信息，导致模型的泛化能力下降；此外，梯度消失或梯度爆炸问题也可能在多模态模型训练中出现，使得训练过程难以收敛；同时，多模态模型的决策过程往往难以解释，由于多种模态数据在模型内部经过复杂的变换和融合，很难确定模型在做某个决策时各个模态数据以及模型内部各层的具体贡

献,这在一些对决策可解释性要求较高的领域,如医疗、金融等,严重限制了多模态人工智能系统的应用^[2]。

(五) 数据稀缺与标注困难

在许多实际应用场景中,获取带有高质量标注的多模态数据十分困难,尤其是在医学、法律等专业领域,数据采集受到严格的伦理和法律限制,数据量也相对有限,同时多模态数据的标注需要跨学科的专业知识,既要理解不同模态数据的含义,又要准确标注它们之间的关联,比如在医学影像与病历文本的多模态数据标注中,就需要医学专业人员同时对影像中的病灶和病历中的症状、诊断信息进行准确标注,这一过程不仅耗时费力,还容易出错;而数据的稀缺性和标注的困难性,会导致多模态模型在训练时缺乏足够的有效数据,难以学习到全面准确的多模态特征表示和融合模式,进而影响模型的性能。

二、应对策略

(一) 统一特征空间构建

为解决数据异构性问题,可以利用深度神经网络学习跨模态映射函数,将不同模态的数据投影到统一的特征空间中。这种方法的核心在于通过对比学习(Contrastive Learning)机制训练模态特定的编码器,使语义相似的样本在嵌入空间中靠近,而不相似的样本远离。例如,在图文检索任务中,通过对比学习训练图像编码器(如基于ViT的视觉Transformer)和文本编码器(如基于BERT的文本Transformer),使相似语义的图像和文本在特征空间中的表示向量距离相近。CLIP(Contrastive Language-Image Pretraining)模型就是一个典型例子,它通过大规模图文对进行预训练,构建了能够对齐视觉和语言特征的统一嵌入空间。

CLIP的核心是通过对比损失函数(Contrastive Loss)实现跨模态特征对齐,该损失函数基于InfoNCE(Noise-Contrastive Estimation)形式,旨在最大化批次中真实匹配对(positive pairs)的相似度,同时最小化非匹配对(negative pairs)的相似度。假设有一个批次包含 N 个图像-文本对,其中 I_i 表示第 i 个图像的嵌入向量, T_i 表示对应的文本嵌入向量。相似度函数 sim 通常采用余弦相似度(Cosine Similarity),定义为:

$$\text{sim}(I_i, T_j) = \frac{I_i \cdot T_j}{\|I_i\| \|T_j\|}$$

其中 $\cdot$ 表示向量点积, $\|\cdot\|$ 表示 L_2 范数。这确保了相似度在 $[-1, 1]$ 范围内规范化,便于比较。

CLIP的损失函数是双向对称的(symmetric),包括图像到文本(image-to-text)和文本到图像(text-to-image)两个方向的对比损失,整体损失为二者的平均:

图像到文本损失:

$$L_{IT} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(I_i, T_i) / \tau)}{\sum_{j=1}^N \exp(\text{sim}(I_i, T_j) / \tau)}$$

总损失:

$$L = \frac{1}{2} (L_{I \rightarrow T} + L_{T \rightarrow I})$$

其中, N 为批次大小, τ 为温度参数(temperature parameter,通常设为0.07左右,用于控制softmax分布的sharpness)。通过最小化该损失,模型学习到将匹配对的相似度分数最大化(接近1),而非匹配对的分数最小化(接近0),从而实现嵌入空间的对齐。在训练过程中,CLIP使用大规模数据集(如400M图文对)进行自监督预训练,优化目标是通过梯度下降更新编码器参数。

在实际应用中,可以基于CLIP模型的思想,针对特定任务(如多模态情感分析)进行微调。具体步骤包括:(1)使用预训练的CLIP编码器提取初始特征;(2)添加任务特定的全连接层;(3)在下游数据集上fine-tune损失函数,以进一步适应域内分布。这种方法不仅解决了异构性,还提升了零样本(zero-shot)泛化能力。

(二) 多模态数据对齐优化

1. 时序对齐算法

针对音频、视频等时序数据的时间轴对齐问题,可以采用动态时间规整(Dynamic Time Warping, DTW)、隐马尔可夫模型(Hidden Markov Model, HMM)等方法。这些算法允许序列在时间维度上非线性伸缩,以最小化对齐后的距离。以语音识别与字幕匹配任务为例,DTW算法通过计算音频特征序列与文本序列的相似度,动态调整时间映射关系,实现精准对齐。其核心是通过递归计算两个序列的最优匹配路径,考虑到时间序列可能的速度变化(如说话快慢不一)。

假设有两个时序序列:音频特征序列 $X = (x_1, x_2, \dots, x_N)$,其中 x_i 为第 i 帧的特征向量(如梅尔频率倒谱系数MFCC,维度通常为13-39维);文本序列 $Y = (y_1, y_2, \dots, y_M)$,其中 y_j 为音素或词级表示(通过音素转换获得)。DTW的目标是找到一条从 $(1, 1)$ 到 (N, M) 的路径 $P = \{(p_1, q_1), \dots, (p_K, q_K)\}$,使得路径上的累积距离最小化,同时满足边界条件(起点 $(1, 1)$,终点 (N, M))、连续性(相邻步长为 $(1, 0)$ 、 $(0, 1)$ 或 $(1, 1)$)。

1)) 和单调性 (路径非递减)。

首先, 构建局部距离矩阵 C , 其中 $C(i, j) = d(x_i, y_j)$, d 通常为欧氏距离 (Euclidean Distance):

$$d(x_i, y_j) = \|x_i - y_j\|_2 = \sqrt{\sum_{k=1}^D (x_{i,k} - y_{j,k})^2}$$

其中 D 为特征维度。为了处理序列长度差异, DTW 使用动态规划构建累积距离矩阵 $D(N \times M \text{ 维})$:

初始化:

$$D(1, 1) = C(1, 1)$$

$$D(i, 1) = D(i-1, 1) + C(i, 1) \text{ for } i = 2 \rightarrow N$$

$$D(1, j) = D(1, j-1) + C(1, j) \text{ for } j = 2 \rightarrow M$$

填充矩阵:

$$D(i, j) = C(i, j) + \min \begin{cases} D(i-1, j) & (\text{垂直移动, 处理Y的插入}) \\ D(i, j-1) & (\text{水平移动, 处理X的插入}) \\ D(i-1, j-1) & (\text{对角移动, 匹配}) \end{cases}$$

最终 DTW 距离为 $D(N, M)$, 表示最小累积成本。最优路径通过回溯 (backtracking) 从 (N, M) 开始, 反向选择最小邻居获得, 用于实际对齐 (如将音频帧映射到字幕时间戳)。

具体实现时, 首先提取音频的 MFCC 特征 (通过短时傅里叶变换和梅尔滤波器组获得), 将文本转换为对应的音素序列 (使用如 CMU Pronouncing Dictionary), 然后利用上述 DTW 公式计算两者之间的最优时间对齐路径, 使得音频和文本在时间上尽可能匹配。为提升效率, 可添加 Sakoe-Chiba 带宽约束 (限制路径偏离对角线不超过 w 步), 减少计算复杂度从 $O(NM)$ 到 $O(Nw)$ 。这种方法在多模态系统中广泛用于同步处理, 确保时序一致性。

2. 语义对齐网络

构建基于注意力机制的语义对齐网络能够自动捕捉不同模态数据间的语义对应关系, 在多模态情感分析中, Transformer 中的多头注意力机制可以聚焦文本中的情感词汇与图像中的表情特征, 建立跨模态的语义关联, 比如可以设计一个多模态 Transformer 模型, 输入文本和图像特征, 通过多头注意力机制让模型自动学习文本和图像中相互关联的部分, 进而实现语义对齐; 在模型训练过程中, 可使用包含丰富情感标注的多模态数据集, 通过监督学习的方式优化模型参数, 使模型能够准确捕捉不同模态数据之间的语义对应关系^[4]。

(三) 计算资源优化策略

1. 模型压缩和知识蒸馏

模型压缩技术能减少模型参数量以降低计算资源消

耗, 比如采用剪枝算法去除模型中不重要的连接或神经元, 运用量化技术将模型参数的数据类型从高精度转换为低精度 (如用 8 位整数代替 32 位浮点数), 而知识蒸馏方法则可将大模型的知识传递给较小的模型, 在保证性能的同时降低计算成本, 具体操作时, 以一个训练好的复杂多模态大模型作为教师模型, 构建一个结构简单的学生模型, 在蒸馏过程中, 让学生模型学习教师模型的输出概率分布, 而非仅仅学习训练数据的标签, 从而使学生模型能在较小的计算资源下尽可能模拟大模型的性能。

2. 多模态模型共享部分网络结构

设计多模态大模型时, 可以让不同模态的网络共享底层表示层, 以此避免每个模态都训练独立的模型, 从而减少计算开销, 比如在一个同时处理图像和文本的多模态模型中, 图像的卷积神经网络和文本的 Transformer 网络可以共享部分底层的特征提取层, 通过联合优化这些共享层, 模型在处理不同模态数据时能够更高效地利用计算资源, 同时也有助于学习到跨模态的通用特征表示, 而在训练过程中, 则可通过多任务学习的方式让模型同时完成与图像和文本相关的任务, 以促进共享层的有效学习。

(四) 提升模型训练稳定性与可解释性

1. 多任务学习

多任务学习让模型在训练时同步优化多个目标函数, 以此避免单一任务可能导致的过拟合问题, 比如在多模态图像描述生成模型中, 可同时设置图像分类任务和图像描述生成任务, 使模型在学习生成图像描述的同时也学习对图像进行分类, 从而从多个角度理解图像和文本信息, 增强泛化能力并提高训练的稳定性, 而在训练过程中, 则需根据不同任务的重要性设置合适的损失权重, 以平衡各个任务对模型训练的影响^[5]。

2. 可解释性模型

利用可解释性技术如注意力机制和可视化工具例如 Grad-CAM, 能够帮助我们理解模型在不同模态融合过程中是如何做出决策的, 其中注意力机制在多模态模型中可以突出显示模型处理不同模态数据时关注的重点区域或关键特征, 通过可视化注意力分布, 我们能直观了解模型是怎样将不同模态信息进行融合和利用的; 而 Grad-CAM 则可以生成与模型决策相关的热力图, 显示图像中对模型决策影响较大的区域, 比如在一个基于多模态数据的疾病诊断模型中, 使用 Grad-CAM 可以展示出图像

中的哪些病变区域以及文本中的哪些症状描述对诊断结果起到了关键作用,进而为医生提供更直观的决策依据。

3. 渐进式训练

通过逐步引入不同模态数据让模型逐渐适应各类数据可提高训练稳定性,训练初期先用单一模态数据对模型预训练使其学习该模态数据的基本特征表示,之后逐步引入其他模态数据与已预训练的模态数据融合训练,比如训练融合图像和文本的多模态情感分析模型时,先用大量图像数据对图像相关网络部分预训练以使其准确提取图像中的情感特征,再引入文本数据与图像数据一起联合训练并调整模型参数,让模型学习到图像和文本之间的情感关联,这种渐进式训练方式能避免模型在训练初期因面对复杂多模态数据而出现训练不稳定的情况。

(五) 解决数据稀缺与标注困难的方法

1. 自监督学习

通过无监督方式利用大规模未标注数据预训练模型是可行的,就像CLIP借助大量未标注的图像和文本对完成训练,进而获得了强大的跨模态表示能力,在实际应用中可采用类似的自监督学习方法,设计合适的自监督任务让模型在未标注的多模态数据上完成预训练,比如针对图像和文本的多模态数据可以设计让模型预测图像和文本是否匹配的任务,通过这种方式模型能够学习到图像和文本之间的语义关联,为后续在少量标注数据上的微调打下良好基础。

2. 生成对抗网络(GAN)与数据增强

利用GAN生成新数据样本或通过数据增强技术扩展现有数据集能够减少对标注数据的需求,在多模态数据中可借助条件GAN根据给定的一种模态数据生成另一种模态数据,比如在语音和文本的多模态场景中就可利用条件GAN根据文本生成对应的语音样本来扩充语音数据,同时对已有的标注数据进行数据增强操作,像对图像进行旋转、缩放、裁剪等变换,对文本进行同义词替

换、随机删除或插入词汇等操作,以此增加数据的多样性并提高模型对不同数据变化的适应性。

3. 迁移学习

为解决标注数据稀缺的问题,可通过迁移学习从其他领域或任务中获取预训练模型,再借助少量标注数据进行微调;在多模态领域,能够利用在大规模通用数据集上预训练的多模态模型,例如在大规模图文数据集上预训练的CLIP模型,针对医学影像与病历文本分析这类特定领域的多模态任务,用该领域少量标注数据对预训练模型进行微调,由于预训练模型已学习到通用的多模态特征表示,所以通过微调能快速适应特定领域的任务,有效减少对大量标注数据的依赖。

结论

多模态人工智能系统中的信息融合虽面临诸多瓶颈,但通过构建统一特征空间、优化数据对齐方法、合理利用计算资源、提升模型训练稳定性与可解释性以及解决数据稀缺与标注困难等一系列有效应对策略,能够逐步突破这些障碍,进而提升多模态人工智能系统的性能和应用效果。

参考文献

- [1] 刘炜.生成式人工智能十大趋势与公共文化机构的应对策略[J].2025(2):25-26.
- [2] 李战子.人工智能时代的多模态话语研究:机遇与路径[J].当代外语研究,2025,25(3):119-120.
- [3] 高跃清.面向多模态数据的认知模型生成及推理应用技术研究[D].北京:北京交通大学,2023(6):54-55.
- [4] 宋奎勇.面向试验数据的多源信息融合方法研究[D].哈尔滨:哈尔滨工程大学,2023(2):23-24.
- [5] 王一帆.基于多模态视频分类任务的模态融合策略研究[J].计算机科学,2024,51(1):8-9.