

智能交互终端的多模态数据融合系统设计

鲍汤胤

杭州立众数字科技有限公司, 浙江杭州, 310000;

摘要: 随着人机交互技术的不断进化, 智能终端逐渐从单一模态的指令响应机制演变为集视觉、听觉、语言、动作等多模态感知与理解为一体的复杂系统。为了实现更加自然、高效、个性化的交互体验, 设计一套完整的多模态数据融合系统成为智能交互终端发展的关键。本研究围绕多模态信息的获取、建模与决策机制, 构建适用于语音识别、图像感知、手势动作与生理信号等多源信息的融合体系, 并以深度神经网络为核心, 提出面向边缘计算的优化部署策略。文章系统分析了传感器协同机制、数据异构处理框架、模态间协同建模与动态适配逻辑, 力求在算力、准确率与交互体验之间实现最优平衡, 推动人机融合新范式的落地应用。

关键词: 智能交互终端; 多模态融合; 深度学习; 边缘计算; 传感器协同; 人机交互

DOI: 10. 64216/3080-1508. 25. 02. 046

引言

在人工智能技术持续演进与人机交互需求日益提升的推动下, 智能交互终端正朝着更智能、更感知、更个性化的方向发展。传统的人机交互主要依赖文本或单一语音命令, 已难以满足复杂场景下对环境理解、用户意图识别与情绪响应等能力的需求。现代智能终端如智能音箱、服务机器人、智能穿戴设备、车载信息系统等, 逐步引入语音、图像、姿态、触觉、眼动、生理信号等多种感知模态, 为系统决策提供更丰富的语义输入。然而, 这些模态数据在空间分布、时间序列、采样频率、语义层级等方面高度异构, 如何实现多模态数据的有效融合与统一决策, 成为设计高效智能交互系统的技术核心。本文以系统工程视角, 系统阐述智能交互终端在多模态融合方面的架构逻辑与技术实现路径, 探索感知协同、语义建模与部署优化之间的协同机制, 为多场景智能终端提供具备可行性、实时性与可拓展性的解决方案。

1 多模态数据感知与同步机制设计

1.1 多类型传感器协同采集结构设计

要让多模态交互系统能够全方位地感知环境以及用户的情况, 关键的第一步就是去搭建一个能把视觉、语音、触觉还有生理信息都囊括进来的传感器感知体系。在硬件架构这块儿呢, 系统会把 RGB 摄像头、深度相机、红外热像仪、麦克风阵列、触觉电极、加速度计、心电/脑电采集模块等等好多不同类型的传感器都集成到一起, 如此一来就形成了一个既能覆盖空间方位, 又能涵盖时间序列的感知网络。每一个传感器在进行物理布置的时候, 都得经过优化来确定摆放的位置点, 要保证它们的视野能够相互补充, 而且

要让噪声干扰尽可能地小, 同时还得依靠时钟同步机制来给它们统一时间戳, 这样才能实现多通道数据在时间上的对齐。在实际进行部署的时候, 通过对 FPGA 或者边缘处理器进行相应的配置, 针对部分模态, 比如说音频, 在本地开展特征提取的工作, 像是提取 MFCC、做频谱分析这些, 这么做一方面能够减少数据传输所需要的带宽, 另一方面也能给后端的融合提供格式统一的接口。凭借着高精度的同步机制以及具有可扩展性的感知架构, 系统就可以从多个不同的角度对用户的姿态、言语、表情、接触行为以及生理状态展开采集的工作, 进而形成作为跨域感知基础的数据流。

1.2 异构模态数据预处理与标准化转换

各类传感器所采集的数据, 在维度、分辨率、频率以及单位这些方面都存在着很明显的差异, 所以在数据要进入主干融合网络之前, 得先做好标准化的预处理工作。系统专门设计了多通道数据预处理引擎, 针对图像数据, 要完成分辨率统一的操作, 还要进行 ROI 区域的提取以及对噪声加以抑制; 对于语音数据, 要开展预加重处理, 做好窗函数分析以及端点检测方面的工作; 而针对动作数据, 则要进行多维向量的重构以及时序的归一化处理。像心电图这类生理信号, 还得进行去基线漂移、小波滤波等去噪方面的处理, 以此来保证后续建模阶段的数据质量能保持一致。为了提高系统整体的响应效率, 预处理模块是通过边缘计算的方式部署在本地嵌入式平台上的, 这样就能实现数据‘边采集边处理’的效果, 并且还能输出格式统一且结构化的中间表示形式, 比如张量矩阵或者特征向量等。经过标准化处理后的数据, 不但方便进行跨模态建模, 而且还能在很大程度上提升后续网络推

理的效率以及准确性。

1.3 时间同步与多模态输入对齐机制

多模态融合系统要达到准确的效果,得依靠各模态输入在时间上实现对齐,并且在语义方面做到同步。不同模态的采样频率不一样,传输路径也存在差异,所以系统得去设计多级的对齐机制,像在硬件层要完成时钟同步,传输层要进行延迟补偿,建模层要开展动态时间规整这些操作。在具体处理的时候,运用多时间窗滑动技术以及交叉模态事件标注的方式,把图像帧、音频波形还有触觉事件相互之间的同步点给对齐匹配好。

同时,借助 Attention 机制给不同模态赋予不一样的时序权重,以此达成动态对齐的目的。除此之外,系统还引入了一种叫“事件驱动触发”的机制,当检测到用户做出意图表达动作,比如张口说话或者手势启动这样的情况时,就会触发多模态采集统一启动起来,以事件作为锚点去构建语义对齐窗口,这样就能有效地提高系统对关键行为的识别准确程度,还能提升系统的时间响应效率。

2 模态融合建模与语义决策机制

2.1 多模态融合网络结构与表示学习机制

多模态融合系统最为关键之处在于要能够把不同感知源的特征有效地进行抽象处理,并且达成语义上的一致性表达。本系统采用的是分层解耦融合这种结构形式,从对特征进行提取一直到做出语义决策,分别设置了与之对应的深度神经网络模块,以此来强化每一个模态的数据表达能力,进而实现最终那种统一的语义表征。在低层进行特征提取这个阶段,图像模态会借助预训练的 CNN 模型,像是 ResNet 或者 MobileNet 这类的,从中提取出局部的纹理以及结构方面的信息;语音模态的输入在经过 MFCC 转换之后,由 LSTM 网络来对语音的时序节奏以及频谱特征加以建模;动作模态则是把加速度计和骨架追踪的数据结合起来,运用 Transformer 框架去捕捉空间动态以及帧间的联系。

到了中层融合阶段,引入了多模态交叉注意力机制,也就是 MCAM,该机制是通过双向注意力权重矩阵的方式,来对各个模态之间相互映射以及语义互补的关系进行建模,并且结合门控单元来动态地调整各个模态的特征表达强度,从而实现信息的重构以及语义的提炼。在高层进行语义建模的时候,是通过共享的嵌入空间,把所有模态的特征都投影到统一的向量域当中,以此来实现像分类、回归或者交互响应等下游任务在语义上保持一致性的处理。在训练阶段,引入了多种增强的机制,其中一方面是通过对比学习来

约束不同模态在同一语义标签之下所表现出来的一致性;另一方面则是通过模态对抗训练去设计那种缺模训练的场景,以此来提高模型在数据缺失、出现异常输入等边缘状态下的容错能力以及稳定性。该网络结构不光是在交互终端进行部署的时候具备高效性以及扩展性,而且还为后续开展个性化建模以及适应语境等方面提供了极为强大的表示基础。

2.2 情绪识别与意图感知的多模态协同机制

要想让智能交互系统达成自然交互以及提供人性化服务,关键就在于理解用户的意图和情绪。本系统是在融合网络这个基础之上构建起了一个能够同时开展情绪识别任务以及意图感知任务的协同模块,依靠多源感知输入来对用户的状态展开细粒度的建模工作。在情绪识别这个模块当中,系统着重对图像还有语音这两种模态加以处理。针对图像模态而言,是凭借表情编码器来识别出面部关键点所发生的变化;而对于语音模态来说,则是要提取诸如语调、音高以及语速等这些时序方面的特征,并且还要融合像眼动轨迹、体态变化等这样的辅助信息,从而构建出高维的情绪特征向量。整个情绪识别的过程是运用时序卷积网络(TCN)来提取出短时情绪的变化趋向,再和多层 Bi-LSTM 相结合以便开展长期依赖的建模,进而能够准确无误地识别出喜悦、悲伤、厌烦、惊讶、愤怒等各种各样的情绪状态。在意图识别模块这边,系统会收集用户的语言输入(也就是自然语言命令)、动作姿态(比如指向、挥手这类动作)、触控行为等诸多信息,将它们融合在一块儿构建出语义行为图谱。通过使用 Bi-GRU 并结合时间序列建模的方式来预测用户接下来即将要执行的操作,再配合规则型意图归纳引擎以及对历史行为轨迹所做的分析,如此一来系统就能够生成像是推荐回应、指令预判或者交互节奏调节等这样的多轮交互策略。

在协同建模机制的作用之下,情绪识别的结果和意图预测的结果会相互交叉,进而成为上下文的影响因子。当用户处于愤怒情绪的时候,就会对指令请求的置信评估进行下调;又或者在用户情绪不太稳定的时候,就会切换交互语调以及输出内容的风格,以此来实现情绪感知和意图理解之间的联动调节。该模块在医疗陪护、智能客服、车载交互、家庭助理等诸多场景当中都有着广泛的适用性,这就使得系统不但能够“听懂”用户的需求,而且还能够“体察”到用户情绪的变化,最终达成真正意义上的智能应对。

2.3 多模态不确定性处理与置信度建模机制

在实际的应用场景当中,多模态输入常常会出现模态有所缺失、数据不够清晰或者感知发生错误等一

系列的问题,所以有必要引入不确定性建模这样的机制来提升系统的鲁棒性。系统引入了贝叶斯推理框架以及模态置信度建模网络,针对每个模态的识别可信度展开动态的评估。在融合层添加 Softmax-Temperature 调节模块以及模态 Dropout 机制,使得系统能够在低置信模态里主动地弱化其自身的贡献,进而防止错误特征对整体判断形成干扰。除此之外,平台还引入了模态纠错机制,当图像和语音出现冲突的时候,借助知识图谱来辅助推理,以此选出更为合理的输出结果。整个的设计确保了系统即便处在信息不完备的状态之下,依旧拥有稳定做出决策的能力,这是达成多模态融合工程可用性以及促使产品落地极为关键的保障。

3 系统部署优化与边缘协同实现路径

3.1 轻量化模型压缩与边缘推理部署机制

智能交互终端往往会被普遍安置在资源方面存在限制的边缘环境当中,所以得把融合网络朝着轻量化以及高效的方向去进行部署操作。系统运用了像是网络剪枝、权重共享、量化压缩还有知识蒸馏等诸多的模型压缩技术,以此来把那种大型的多模态模型压缩成能够被部署到NPU或者TPU这类边缘设备之上的格式。就部署架构来讲,构建起了以ONNX作为中间表示形式,同时将TensorRT当作加速引擎的推理框架,进而达成跨平台的算子调度以及硬件层面的优化。在关键的交互环节里面,采用了小模型开展局部推理再加上大模型进行远程请求的分层模式,这样就能够实现‘对延迟敏感的任务在边缘处完成,而复杂分析任务则通过云端来协同’的混合处理策略,从而提升系统的实时性能以及能耗比。

3.2 跨设备协同与分布式模态感知机制

在多终端协同这样的场景之中,系统得要达成多个智能终端相互间的模态协同以及任务分摊这两方面的要求。平台着手去构建模态任务调度器,把语音交互方面的事宜交给智能音箱来完成,视觉识别这块儿呢,则安排由摄像头端去处理,至于生理信号则交由可穿戴设备来进行感知。通过局域网以及Mesh网络,实现数据能够统一汇聚起来并且完成融合推理的操作。为了确保协同过程当中的低延迟情况以及数据的一致性能得以保证,系统依据边缘计算节点来构建模态同步协议还有边缘分布式数据库,以此来提高模态组合方面的能力以及对不同场景的适应能力。尤其是处在智慧家居、智能车载等存在多个节点交互的

环境里面的时候,该机制在很大程度上提升了系统所具备的分布式智能的水平。

3.3 实时响应优化与任务卸载机制设计

鉴于不同交互任务对于系统实时性的要求存在着颇为显著的差异,平台特意引入了那种基于优先级而构建起来的任务调度模型,依照实际需求去分配相应的计算资源。当系统处于负载的高峰期之时,平台能够把一部分并非关键模态的任务采取本地缓存的方式来处理,或者也可以进行异步处理,就像是那些有所延迟的手势分析任务、相对次要的生理监测数据等等,借助智能任务卸载机制把这些任务传输到远端的服务器那里,进而开展异步建模以及知识的回传操作。系统对任务优先队列以及权重参数都做出了设定,再结合用户的使用偏好以及网络的实际状态,以一种动态的方式去调整卸载的策略,从而达成算力负载能够实现动态均衡的效果,并且还能让响应延时达到最小化的程度。这样的一种机制,在确保核心交互链路可以保持稳定的同时,还能够提升系统针对复杂交互环境所具备的资源调度方面的能力。

4 结语

多模态数据融合是实现智能交互终端真正“理解人、感知意、主动反应”的核心路径。本文从感知协同、建模机制到系统部署三个层面,系统阐述了智能交互终端多模态融合系统的设计方法与技术路径。通过合理的传感器布局、深度融合网络结构与边缘协同部署策略,可实现多模态交互系统在复杂现实环境下的高效运行与持续进化。未来,随着计算能力提升与感知算法发展,多模态融合系统将更加注重个性化、情境感知与人文交互,推动人机共生的智能时代加速到来。

参考文献

- [1]陶建华,巫英才,喻纯,等.多模态人机交互综述[J].中国图象图形学报.2022,27(6).
- [2]薛志荣.AI改变设计:人工智能时代的设计师生存手册[M].清华大学出版社,2018.
- [3]刘文杰.城市交通大数据可视化框架及实现[J].科技创新导报,2017,14(36):121-122.
- [4]陈为,朱标,张宏鑫.BN-Mapping:基于贝叶斯网络的地理空间数据可视分析[J].计算机学报,2016,39(7):12811293
- [5]李一凡,杨帆,余正红.城市交通时空数据可视化研究[J].软件导刊,2021,20(5):139-147.