(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114936552 B
(45)授权公告日 2025.06.13

(21)申请号 202210624963.3

(22)申请日 2022.06.02

(65)同一申请的已公布的文献号
申请公布号 CN 114936552 A

(43)申请公布日 2022.08.23

(73)专利权人 杭州电子科技大学
地址 310018 浙江省杭州市下沙高教园区2号大街

(72)发明人 孔万增 郭继伟 唐佳佳 戴玮辰 刘栋军

(74)专利代理机构 杭州君度专利代理事务所(特殊普通合伙) 33240
专利代理师 陈炜

(51)Int.Cl.
G06F 40/284(2020.01)
G06F 40/30(2020.01)
G06F 18/2431(2023.01)
G06F 18/213(2023.01)
G06F 18/25(2023.01)
G06N 3/045(2023.01)
G06N 3/0442(2023.01)

(56)对比文件
Jiwei Guo等.Dynamically AdjustWord Representations Using Unaligned Multimodal Information.《View Web of Science ResearcherID and ORCID》.2022,第3394-3402页.

审查员 徐生芹

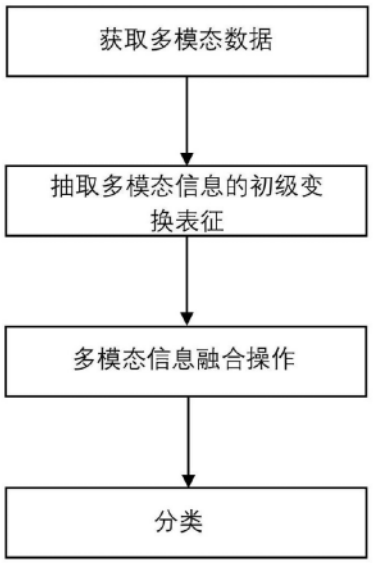
权利要求书2页 说明书7页 附图2页

(54)发明名称

未对齐行为信息动态调整单词表示的多模态情感识别方法

(57)摘要

本发明公开一种未对齐行为信息动态调整单词表示的多模态情感识别方法。本发明利用跨模态注意力机制,挖掘与文本模态相关的行为信息(由视觉是听觉模态组成),然后利用行为信息来动态的修改文本模态中的单词在语义空间中的位置,从而得到经过多模态信息调整后的单词表示。同时,跨模态注意力机制能够在长距离范围内关注到与文本模态相关的行为信息,因此能够很好的解决多模态学习中存在的固有问题—各个模态信息之间的频率不匹配。其次,在此基础上构建若干个多模态Transformer层,能够进一步挖掘经过多模态信息调整后的单词表示在上下文环境中的高级特征信息,是对当前情感识别领域的多模态融合框架的有效补充。



CN 114936552 B

1.未对齐行为信息动态调整单词表示的多模态情感识别方法,其特征在于:包括以下步骤:

步骤1、数据采集;

获取含有不同情感类别下采集的多模态数据集;

步骤2、多模态信息数据预处理;

分别将文本模态、视觉模态、听觉模态数据转化为初级表征,并对听觉和视觉模态数据进行预融合操作,降低听觉和视觉模态数据的时域维度尺寸以及特征向量长度大小;

采用长短期记忆网络抽取视觉和听觉数据的初级特征如下:

$$F_m = \text{sLSTM}\left(I_m; \theta_m^{lstm}\right) \in R^{T_m \times d_m} \qquad \text{公式(1)}$$

其中,F_m为视觉或听觉数据的初级特征,$m \in \{v, a\}$,$F_m \in \mathbb{R}^{T_m \times d_m}$为模态m的初级表征;v、a分别表示视觉、听觉模态;I_m为模态m的原始数据;θ_m^{lstm}为模态m的重矩阵;T_m为时域维度的尺寸;d_m为每一个时刻的特征向量的长度;

听觉或视觉模态数据的预融合的结果$X_{\{m\}}$的表达式如下:

$$X_{\{m\}} = \text{Conv 2D}\left(\{m\}, k_{\{m\}}\right) \in \mathbb{R}^{T_{\{m\}} \times d_m} \qquad \text{公式(2)}$$

其中,{m}为模态m的初级表征;T_m为时域维度的尺寸,d_m为每一个时刻的特征向量的长度;$k_{\{m\}}$为模态m的卷积核的大小;

步骤3、跨超模态融合;

3-1.获取超模态信息

将经过预融合操作的视觉和听觉模态的初级表征在时域维度上拼接在一起,得到超模态信息X_β;

3-2.动态调整单词表示;

将超模态信息分别经过两个线性转换网络,得到关键矩阵K_β以及实值矩阵V_β;将文本模态信息经过一个线性转换网络,得到对应的查询矩阵Q_l;

基于查询矩阵Q_l和关键矩阵K_β计算得到行为信息在文本模态中的注意力因子矩阵e如下:

$$a = \frac{Q_l K_\beta^T}{\sqrt{d_k}}$$

$$e = \text{softmax}(a) \qquad \text{公式(6)}$$

其中,a为未归一化的注意力因子矩阵;d_k为查询矩阵Q_l的特征长度;

提取超模态信息中与文本相关的信息H如下:

$$H = eV_\beta \qquad \text{公式(7)}$$

获取融入了未对齐行为信息的文本信息;

利用上述得到的超模态信息中与文本相关的信息H动态调整文本模态中的每一个单词表示如下:

$$\overline{X}_l = X_l + \alpha H$$

$$\alpha = min\left(\frac{\|X_l\|_2}{\|H\|_2}\lambda, 1\right) \qquad \text{公式(8)}$$

其中，$\overline{X}_l$表示融入了超模态信息的文本信息；X_1表示文本模态的初始表征；α为比例系数；λ为预设的超参数；

以文本信息$\overline{X}_l$输入情感识别模型中进行训练；

步骤四、情感识别输出

采集被测对象的多模态数据送入步骤三获取的情感识别模型，识别被测对象的情感类别。

2.根据权利要求1所述的未对齐行为信息动态调整单词表示的多模态情感识别方法，其特征在于：所述的情感类别包括积极情绪和消极情绪。

3.根据权利要求1所述的未对齐行为信息动态调整单词表示的多模态情感识别方法，其特征在于：步骤2中，通过预训练语言模型将文本信息经过文本编码转化为词嵌入方式的初级表征。

4.根据权利要求1所述的未对齐行为信息动态调整单词表示的多模态情感识别方法，其特征在于：关键矩阵K_β以及实值矩阵V_β的表达式如下如下：

$$K_\beta = X_\beta W_{K_\beta}$$

$$W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}$$

$$V_\beta = X_\beta W_{V_\beta}$$

$$W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v} \qquad \text{公式（4）}$$

其中，W_{K_β}，W_{V_β}分别是矩阵K_β，V_β的线性网络的权重矩阵；d_β，d_k，d_v分别为超模态信息、关键矩阵、实值矩阵的特征向量长度。

5.根据权利要求1所述的未对齐行为信息动态调整单词表示的多模态情感识别方法，其特征在于：查询矩阵Q_1的表达式如下：

$$Q_l = X_l W_{Q_l}$$

$$W_{Q_l} \in \mathbb{R}^{d_l \times d_k} \qquad \text{公式（5）}$$

其中，X_1为文本模态信息，W_{Q_l}是查询矩阵的权重矩阵；d_1和d_k分别为文本模态和查询矩阵的特征向量长度。

6.根据权利要求1所述的未对齐行为信息动态调整单词表示的多模态情感识别方法，其特征在于：所述的情感识别模型采用BERT模型。

7.一种情感识别系统，包括处理器和存储器；其特征在于：所述存储器存储有能够被所述处理器执行的机器可执行指令，所述处理器执行所述机器可执行指令以实现如权利要求1-6中任意一项所述的多模态情感识别方法；机器可执行指令包括数据采集模块、数据预处理模块、跨超模态融合和情感识别输出模块。

8.一种机器可读存储介质；其特征在于：该机器可读存储介质存储有机器可执行指令，该机器可执行指令在被处理器调用和执行时，机器可执行指令促使处理器实现如权利要求1-6中任意一项所述的多模态情感识别方法。

未对齐行为信息动态调整单词表示的多模态情感识别方法

技术领域

[0001] 本发明属于自然语言处理、视觉、语音交叉领域内的多模态情感识别领域,具体涉及一种使用未对齐行为信息动态调整单词表示的多模态情感识别方法,利用基于跨模态注意力机制的融合网络技术,对由视觉和听觉组成的行为信息和文本模态信息在整个话语范围尺度内进行长时融合,动态地转移单词在语义空间中的位置从而判断被试情感状态。

背景技术

[0002] 情感分析领域通常包含文本模态,视频模态以及语音模态等数据信息。在以往的研究中,验证了这些单模态数据中包含着与情感状态相关的判别信息。同时,研究发现,这些单模态数据之间存在的一致性和互补性能够有效解释多模态数据内部的关联表征,并且能够进一步增强模型表达能力及稳定性,提升情感任务分析性能。

[0003] 现有的基于调整单词表示的多模态融合模型,由于能够对细粒度多模态信息数据进行有效建模,从而能够在一定程度上减轻使用平均的策略而导致忽略局部模态内部的复杂交互信息带来的影响,因此引起了广泛关注。具体操作为:在将多模态融合的过程中,首先分别对视觉与文本之间两个模态进行融合、对听觉与文本两个模态进行融合,然后将两种融合后的信息继续融合,从而得到包含所有模态的融合信息。但是当多模态的数量超过两个时,需要进行多次的双模态融合操作后才能得到包含所有模态的融合信息。这种双向的融合策略将会导致模型保留大量的原始参数,极大影响模型的性能表现。此外,现有调整单词表示的网络通常利用手动对齐的多模态序列数据来动态调整单词在语义空间中表示。由于每种模态的采样率不同,因此采集到的多模态序列数据通常是非对齐的。在对齐的行为信息中调整单词表示,首先要将行为信息与文本模态进行对齐操作,使三种模态信息在时间维度上保持一致。然而,在深度学习任务中,标注这一操作需要耗费大量的人力物力成本,因此利用未对齐的行为信息相比对齐的行为信息去动态调整单词表示是具有现实意义的方法。

发明内容

[0004] 本发明的目的是针对现有技术的不足,提出一种未对齐的行为信息动态调整单词表示的多模态情感分类方法。

[0005] 第一方面,本发明提供一种未对齐行为信息动态调整单词表示的多模态情感识别方法,其包括以下步骤:

[0006] 步骤1、数据采集。

[0007] 获取含有不同情感类别下采集的多模态数据集。

[0008] 步骤2、多模态信息数据预处理。

[0009] 分别将文本模态、视觉模态、听觉模态数据转化为初级表征,并对听觉和视觉模态数据进行预融合操作,降低听觉和视觉模态数据的时域维度尺寸以及特征向量长度大小。

[0010] 步骤3、跨超模态融合。

[0011] 3-1.获取超模态信息

[0012] 将经过预融合操作的视觉和听觉模态的初级表征在时域维度上拼接在一起,得到超模态信息X_β。

[0013] 3-2.动态调整单词表示。

[0014] 将超模态信息分别经过两个线性转换网络,得到关键矩阵K_β以及实值矩阵V_β;将文本模态信息经过一个线性转换网络,得到对应的查询矩阵Q_1。

[0015] 基于查询矩阵Q_1和关键矩阵K_β计算得到行为信息在文本模态中的注意力因子矩阵e如下:

[0016] $$a=\frac{Q_l K_\beta^T}{\sqrt{d_k}}$$

[0017] e=softmax(a) 公式(6)

[0018] 其中,a为未归一化的注意力因子矩阵;d_k为查询矩阵Q_1的特征长度。

[0019] 提取超模态信息中与文本相关的信息H如下:

[0020] $H=eV_\beta$ 公式(7)

[0021] 获取融入了未对齐行为信息的文本信息;

[0022] 利用上述得到的超模态信息中与文本相关的信息H动态调整文本模态中的每一个单词表示如下:

[0023] $$\overline{X}_l = X_l + \alpha H$$
$$\alpha = min\left(\frac{\|X_l\|_2}{\|H\|_2}\lambda, 1\right)$$ 公式(8)

[0024] 其中,$\overline{X}_l$表示融入了超模态信息的文本信息。X_1表示文本模态的初始表征;α为比例系数;λ为预设的超参数。

[0025] 以文本信息$\overline{X}_l$输入情感识别模型中进行训练。

[0026] 步骤四、情感识别输出

[0027] 采集被测对象的多模态数据送入步骤三获取的情感识别模型,识别被测对象的情感类别。

[0028] 作为优选,所述的情感类别包括积极情绪和消极情绪。

[0029] 作为优选,步骤2中,通过预训练语言模型将文本信息经过文本编码转化为词嵌入方式的初级表征。

[0030] 作为优选,步骤2中,采用长短期记忆网络抽取视觉和听觉数据的初级特征如下:

[0031] $$F_m = \text{sLSTM}\left(I_m; \theta_m^{lstm}\right) \in R^{T_m \times d_m}$$ 公式(1)

[0032] 其中,F_m为视觉或听觉数据的初级特征,$m \in \{v, a\}$,$F_m \in \mathbb{R}^{T_m \times d_m}$为模态m的初级表征;v、a分别表示视觉、听觉模态;I_m为模态m的原始数据;θ_m^{lstm}为模态m的重矩阵;T_m为时域维度的尺寸;d_m为每一个时刻的特征向量的长度。

[0033] 作为优选,步骤2中,听觉或视觉模态数据的预融合的结果$X_{\{m\}}$的表达式如下:

[0034] $$X_{\{m\}} = \text{Conv 2D}\left(\{m\}, k_{\{m\}}\right) \in \mathbb{R}^{T_{\{m\}} \times d_m}$$ 公式(2)

[0035] 其中,{m}为模态m的初级表征;T_m为时域维度的尺寸,d_m为每一个时刻的特征向量

的长度;$k_{\{m\}}$为模态m的卷积核的大小。

[0036] 作为优选,关键矩阵K_β以及实值矩阵V_β的表达式如下如下:

[0037] $$K_\beta = X_\beta W_{K_\beta}$$

[0038] $$W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}$$

[0039] $$V_\beta = X_\beta W_{V_\beta}$$

[0040] $$W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$$ 公式(4)

[0041] 其中,W_{K_β}, W_{V_β}分别是矩阵K_β,V_β的线性网络的权重矩阵;d_β,d_k,d_v分别为超模态信息、关键矩阵、实值矩阵的特征向量长度。

[0042] 作为优选,查询矩阵Q_l的表达式如下:

[0043] $$Q_l = X_l W_{Q_l}$$

[0044] $$W_{Q_l} \in \mathbb{R}^{d_l \times d_k}$$ 公式(5)

[0045] 其中,X_l为文本模态信息,W_{Q_l}是查询矩阵的权重矩阵;d_l和d_k分别为文本模态和查询矩阵的特征向量长度。

[0046] 作为优选,所述的情感识别模型采用BERT模型(Bidirectional Encoder Representation from Transformers)。

[0047] 第二方面,本发明提供一种情感识别系统,其包括处理器和存储器。所述存储器存储有能够被所述处理器执行的机器可执行指令,所述处理器执行所述机器可执行指令以实现前述的多模态情感识别方法。机器可执行指令包括数据采集模块、数据预处理模块、跨超模态融合和情感识别输出模块。

[0048] 第三方面,本发明提供一种机器可读存储介质;该机器可读存储介质存储有机器可执行指令,该机器可执行指令在被处理器调用和执行时,机器可执行指令促使处理器实现前述的多模态情感识别方法。

[0049] 本发明的有益效果是:

[0050] 本发明结合跨模态注意力机制,利用未对齐的行为信息动态调整文本模态中的单词表示,挖掘非文本模态对文本模态之间的长时交互的模态融合信息。此外,跨模态注意力机制能够同时对多个模态信息进行建模操作,因此能很好地应对多模态学习中存在的固有问题——多个模态不能同时进行交互。紧接着,在此基础上构建了多模态Transformer框架,将经过行为信息动态调整后的单词表示送入其中,进一步进行高层次的多模态融合,是对当前情感识别领域的多模态融合框架的有效补充。

附图说明

[0051] 图1为本发明的流程图;

[0052] 图2为本发明中动态调整单词网络的示意图;

[0053] 图3为三模态融合示意图。

具体实施方式

[0054] 下面结合附图,对本发明方法做详细描述。

[0055] 如图1所示,一种未对齐行为信息动态调整单词表示的多模态情感识别方法,包括以下步骤:

[0056] 步骤1、获取多模态信息数据

[0057] 在被试执行特定情感任务的过程中,记录被试的文本模态数据、语音模态数据以及视频模态数据,作为多模态数据集。特定情感任务包括积极情绪和消极情绪,具体可以细分为非常消极,消极,弱消极,中性,弱积极,积极,非常积极。

[0058] 步骤2、多模态信息数据预处理

[0059] 多模态数据是在特征层面上进行多模态融合操作;对于文本模态,采用预训练语言模型,将原始的文本信息经过文本编码(text encoder)转化为词嵌入(Embedding)方式的初级表征。

[0060] 对于听觉和视觉模态,采用长短期记忆网络抽取视觉和听觉数据的初级特征表示;

[0061] $$F_m = \mathrm{sLSTM}\left(I_m; \theta_m^{lstm}\right) \in R^{\mathrm{T_m}\times \mathrm{d_m}} \qquad \text{公式(1)}$$

[0062] 其中,F_m为视觉或听觉数据的初级特征,$m \in \{v, a\}$,$F_m \in \mathbb{R}^{T_m \times d_m}$为模态m的初级表征;v、a分别表示视觉、听觉模态;I_m为模态m的原始数据;θ_m^{lstm}为模态m的重矩阵;T_m为时域维度的尺寸;d_m为每一个时刻的特征向量的长度;由于模态采样率的标准不同,非文本模态(视觉和听觉模态)的时域维度尺寸通常比文本模态时域维度的尺寸要大得多,不利于多模态融合操作。为此,针对听觉和视觉模态进行预融合操作,降低其时域维度尺寸以及特征向量长度大小;

[0063] $$X_{\{m\}} = \mathrm{Conv\ 2D}\left(\{m\}, k_{\{m\}}\right) \in \mathbb{R}^{T_{\{m\}} \times d_m} \qquad \text{公式(2)}$$

[0064] 其中,$X_{\{m\}} \in \mathbb{R}^{T_{\{m\}} \times d}$为模态m预融合的结果;$T_m$为时域维度的尺寸,$d_m$为每一个时刻的特征向量的长度;$k_{\{m\}}$为模态m的卷积核的大小。Conv 2D(•)为二维卷积处理。

[0065] 步骤3、基于跨超模态融合方法,利用未对齐的视觉和听觉模态信息,动态调整文本模态在语义空间中的表示。该方法包括获取超模态信息和动态调整单词表示两个任务;

[0066] 3-1.获取超模态信息

[0067] 获取超模态信息的学习过程,将经过预融合操作的未对齐视觉和听觉模态的初级表征在时域维度上拼接在一起,得到超模态信息。这种超模态信息包含了影响文本表示的全部信息。包含视觉和听觉模态的超模态信息的表达式如下:

[0068] $$X_\beta = v \oplus a \qquad \text{公式(3)}$$

[0069] 其中,X_β表示获得到的超模态信息,v表示视觉模态信息,a表示听觉模态信息,$\oplus$表示拼接操作。

[0070] 3-2.动态调整单词表示。动态调整单词表示的学习过程,对于文本模态中的每一个单词表示,利用前述得到的超模态信息,在整个话语尺度范围内动态调整文本模态的每一个单词表示,将视觉和听觉模态组成的超模态信息融入到文本表示中,从而完成多模态融合,具体过程如下:

[0071] 将超模态信息分别经过两个线性转换网络,得到对应的关键矩阵K_β以及实值矩阵

V_β,表示如下:

[0072] $$K_\beta = X_\beta W_{K_\beta}$$

[0073] $$W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}$$

[0074] $$V_\beta = X_\beta W_{V_\beta}$$

[0075] $$W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$$ 公式(4)

[0076] 其中,W_{K_β}, W_{V_β}分别是矩阵K_β,V_β的线性网络的权重矩阵;d_β,d_k,d_v分别为超模态信息、关键矩阵、实值矩阵的特征向量长度。

[0077] 将文本模态信息经过一个线性转换网络,得到对应的查询矩阵Q_l,表示如下:

[0078] $$Q_l = X_l W_{Q_l}$$

[0079] $$W_{Q_l} \in \mathbb{R}^{d_l \times d_k}$$ 公式(5)

[0080] 其中,X_l为文本模态信息,W_{Q_l}是查询矩阵Q_l的权重矩阵;d_l和d_k分别为文本模态和查询矩阵的特征向量长度。

[0081] 利用跨模态注意力机制,将超模态信息融入到文本模态中,利用行为信息来动态调整单词在语义空间中表示,具体如下:

[0082] 对于跨模态注意力机制,基于查询矩阵Q_l和关键矩阵K_β计算得到行为信息在文本模态中的注意力因子矩阵e如下:

[0083] $$a = \frac{Q_l K_\beta^T}{\sqrt{d_k}} = \frac{X_l W_{Q_l} W_{K_\beta}^T X_\beta^T}{\sqrt{d_k}} \in \mathbb{R}^{T_l \times T_\beta}$$

[0084] $$e = \text{softmax}(a) = \text{softmax}\left(\frac{X_l W_{Q_l} W_{K_\beta}^T X_\beta^T}{\sqrt{d_k}}\right)$$ 公式(6)

[0085] 其中,a为未归一化的注意力因子矩阵;d_k为查询矩阵Q_l的特征长度。

[0086] 根据注意力因子矩阵和实值矩阵相作用,得到超模态信息与文本信息在时域上的长时相关性;

[0087] $$H = eV_\beta = \text{softmax}(a)V_\beta = \text{softmax}\left(\frac{Q_l K_\beta^T}{\sqrt{d_k}}\right)V_\beta$$ 公式(7)

[0088] 其中,H表示超模态信息中与文本相关的信息。

[0089] 利用上述得到的超模态信息中与文本相关的信息H动态调整文本模态中的每一个单词表示,表示如下:

[0090] $$\overline{X}_l = X_l + \alpha H$$
$$\alpha = min\left(\frac{\|X_l\|_2}{\|H\|_2}\lambda, 1\right)$$ 公式(8)

[0091] 其中,X_l表示未经调整的文本模态信息,$\overline{X}_l$表示融入了未对齐行为信息的文本信息。α为比例系数;λ为预设的超参数;$||\cdot||_2$为二范数运算。

[0092] 融入了未对齐行为信息的文本信息$\overline{X}_l$中,添加了视频以及音频模态信息,极大地

补充了单一文本模态信息表达能力的局限性。我们在每一个文本模态前面添加一个特殊的标记(CLS)用做多模态情感分类的标签。原始的文本模态信息经过上述操作后都会得到一个新的文本模态表示向量，将汇聚了多模态信息的$\overline{X}_l = \{l_{CLS}, l'_1, l'_2, \cdots, l'_{n+1}\}$送入BERT的Transformers层中继续训练，得到情感识别模型，用于下游的情感分类任务。训练的损失函数为$\text{loss}\ (\hat{y}, y) = \frac{1}{n}\sum(\hat{y}_i - y_i)^2$

[0093] 步骤四、同时提取被测对象的文本模态、视觉模态和听觉模态信息，并输入情感识别模型，获取被测对象所处的情绪类别。

[0094] 图2为使用未对齐的多模态信息动态调整单词表示操作流程图。图3为三个模态A、V以及T的多模态融合流程图。

[0095] 使用本发明与多种现有的多模态融合方法同时在两个公开的多模态情感数据库CMU-MOSI、CMU-MOSEI上进行情感状态判别任务，每种数据集都有对齐和未对齐两种格式的数据，结果如表1、2所示；表中结果为平均绝对误差MAE、相关系数Corr、情感二分类任务对应的精确度Acc-2、F1分数F1-Score以及情感七分类任务对应的精度Acc-7。可以看出，与表现出优异水平的现有多模态融合框架相比，本发明的五个评价指标均优于现有融合模型，证明了本发明所提出方法的有效性。

[0096] 表1.结果对比表

[0097]

Metrics	MAE	Corr	Acc-2	F1-Score	Acc-7
(Word Aligned) CMU-MOSI Sentiment					
MFN†	0.965	0.632	77.4/-	77.3/-	34.1
RAVEN†	0.915	0.691	78.0/-	76.6/-	33.2
MulT	0.871	0.698	-/83.0	-/82.8	40.0
MFM †	0.877	0.706	-/81.7	-/81.6	35.4
ICCN†	0.860	0.710	-/83.0	-/83.0	39.0
MISA	0.783	0.761	81.8/83.4	81.7/83.6	42.3
MAG*	0.727	0.781	82.37/84.43	82.50/84.61	43.62
(Unaligned) CMU-MOSI Sentiment					
TFN†	0.901	0.698	-/80.8	-/80.7	34.9
MulT	0.889	0.686	-/81.1	-/81.0	39.1
MTAG	0.866	0.722	-/82.3	-/82.1	38.9
Self-MM*	0.712	0.795	82.54/84.77	82.68/84.91	45.79
MMIM	0.700	0.800	84.14/86.06	84.00/85.98	46.65
CHFN	**0.689**	**0.809**	**84.3/86.4**	**84.2/86.2**	**48.6**

[0098] 表2.结果对比表

[0099]

Metrics	MAE	Corr	Acc-2	F1-Score	Acc-7
(Word Aligned) CMU-MOSEI Sentiment					
MFN†	-	-	76.0/-	76.0/-	-
RAVEN†	0.614	0.662	79.1/-	79.5/-	50.0
MulT	0.580	0.703	-/82.5	-/82.3	51.8
MFM†	0.568	0.717	-/84.4	-/84.3	51.3
ICCN†	0.565	0.713	-/84.2	-/84.2	51.6
MISA	0.555	0.756	83.6/85.5	83.8/85.3	52.2
MAG*	0.543	0.755	82.51/84.82	82.77/84.71	52.67
(Unaligned) CMU-MOSEI Sentiment					
TFN†	0.593	0.700	-/82.5	-/82.1	50.2
MulT	0.591	0.694	-/81.6	-/81.6	50.7
Self-MM*	0.529	0.767	82.68/84.96	82.95/84.93	53.46
MMIM	0.526	0.772	82.24/85.97	82.66/85.94	54.24
CHFN	**0.525**	**0.778**	**83.7/86.2**	**83.9/86.1**	**54.30**

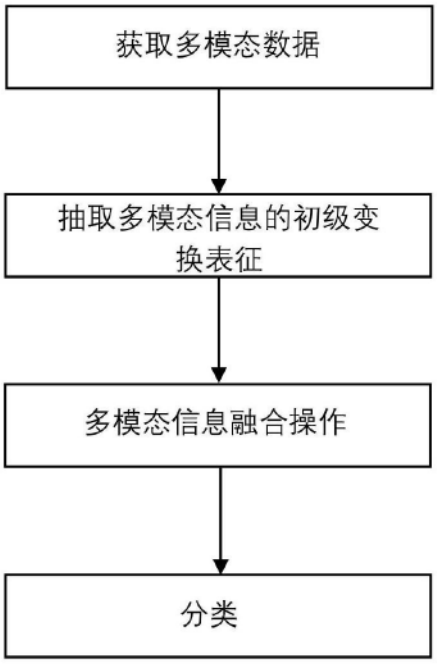

获取多模态数据
抽取多模态信息的初级变换表征
多模态信息融合操作
分类

图1

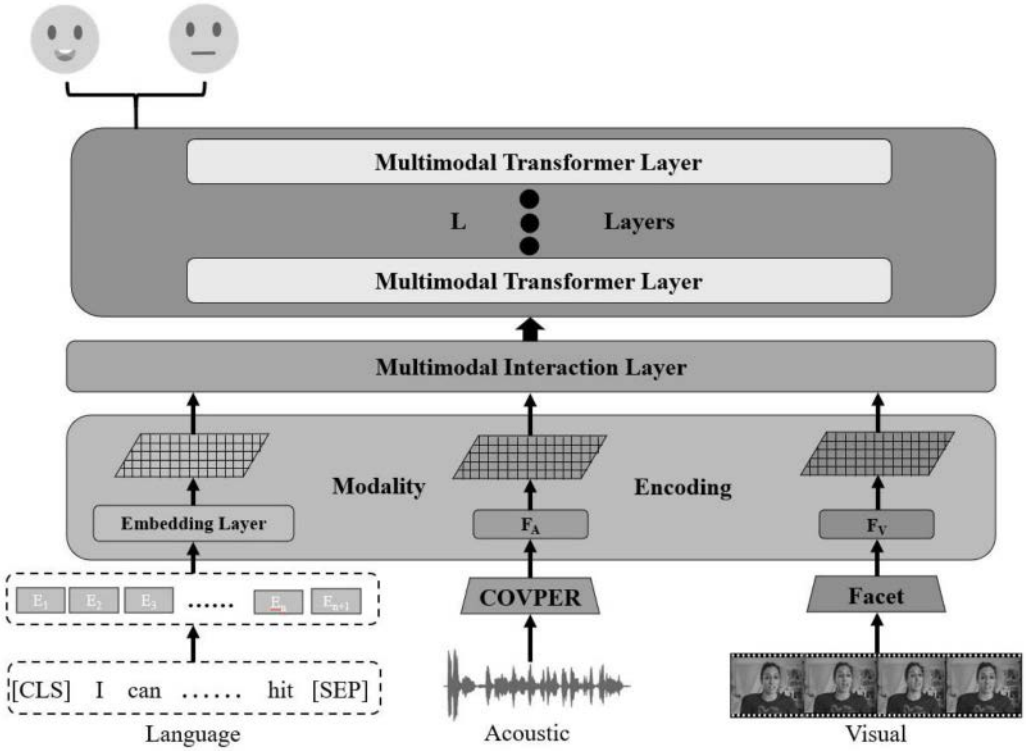

Multimodal Transformer Layer
L
Layers
Multimodal Transformer Layer
Multimodal Interaction Layer
Modality
Encoding
Embedding Layer
F_A
F_V
COVPER
Facet
[CLS] I can ..... hit [SEP]
Language
Acoustic
Visual

图2

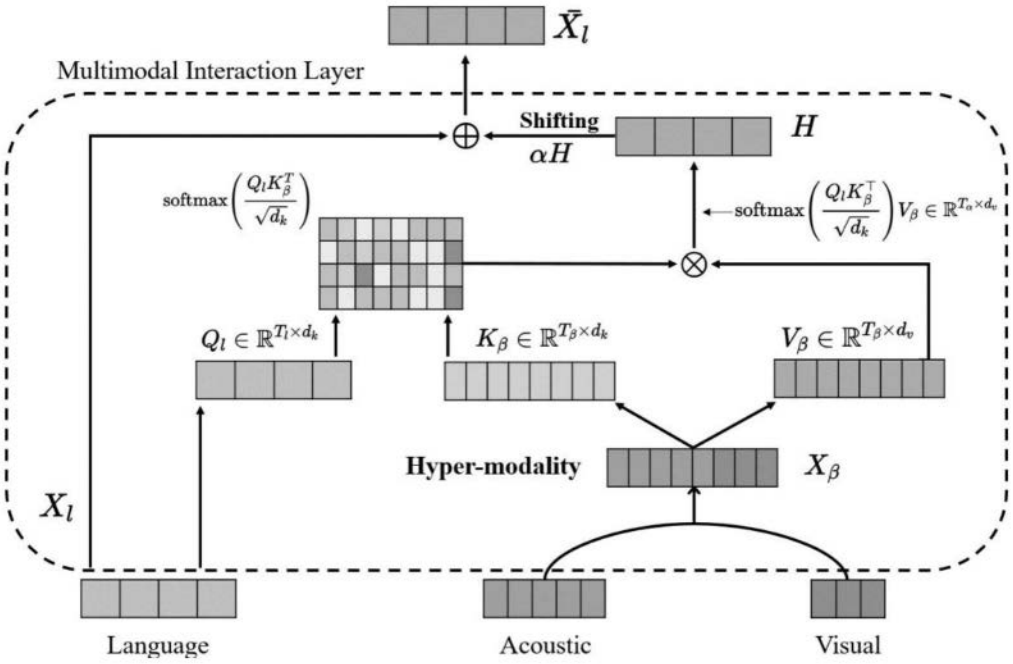

$\bar{X}_l$
Multimodal Interaction Layer
Shifting
αH
H
$\text{softmax}\left(\frac{Q_l K_\beta^T}{\sqrt{d_k}}\right)$
$\text{softmax}\left(\frac{Q_l K_\beta^\top}{\sqrt{d_k}}\right) V_\beta \in \mathbb{R}^{T_\alpha \times d_v}$
$Q_l \in \mathbb{R}^{T_l \times d_k}$
$K_\beta \in \mathbb{R}^{T_\beta \times d_k}$
$V_\beta \in \mathbb{R}^{T_\beta \times d_v}$
Hyper-modality
X_β
X_l
Language
Acoustic
Visual

图3