(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113269277 B
(45)授权公告日 2023.07.25

(21)申请号 202110759965.9
(22)申请日 2021.07.06
(65)同一申请的已公布的文献号
申请公布号 CN 113269277 A
(43)申请公布日 2021.08.17
(66)本国优先权数据
202010733299.7 2020.07.27 CN
(73)专利权人 西北工业大学
地址 710072 陕西省西安市友谊西路127号
(72)发明人 陈海丰 蒋冬梅
(74)专利代理机构 西安凯多思知识产权代理事务所(普通合伙) 61290
专利代理师 云燕春
(51)Int.Cl.
G06F 18/25(2023.01)
G06N 3/0455(2023.01)
G06N 3/0464(2023.01)
G06N 3/049(2023.01)
G06N 3/08(2023.01)

(56)对比文件
CN 110728997 A,2020.01.24
CN 109919205 A,2019.06.21
CN 111368536 A,2020.07.03
CN 110188343 A,2019.08.30
CN 111275085 A,2020.06.12
CN 109543824 A,2019.03.29
US 2004002270 A1,2004.01.01
US 2019130273 A1,2019.05.02
Jian Huang等.MULTIMODAL TRANSFORMER FUSION FOR CONTINUOUS EMOTION RECOGNITION.《 ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)》.2020,
Emre Aksan等.Attention, please: A Spatio-temporal Transformer for 3D Human Motion Prediction.《arXiv:2004.08692v1 [cs.CV] 18 Apr 2020》.2020,
陈珂等.基于情感词典和 Transformer 模型的情感分析算法研究.《南京邮电大学学报(自然科学版)》.2020,第40卷(第1期),

审查员 周亚芳

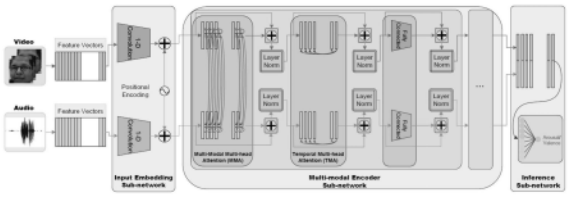
权利要求书1页 说明书5页 附图1页

(54)发明名称

基于Transformer编码器和多头多模态注意力的连续维度情感识别方法

(57)摘要

本发明采用时序深度卷积神经网络(TCN)、自注意力Transformer编码器(Transformer Encoder)以及多模态多头注意力机制(Multi-modal Multi-head Attention),涉及一种从多模态(听觉、视觉)时序信息中对连续维度情感进行估计的模型和识别方法。该方法对不同模态输入的特征,得到不同模态的嵌入特征表达;而后将不同模态的嵌入特征表达作为输入,利用多模态Transformer编码器得到不同模态的高级表征;最后将不同模态的高级特征表达作为输入,计算出每一时刻的情感状态值。本发明更加关注时序上过去某些关键时刻对当前情绪状态的影响,排除长远情感信息带来的干扰,使得模型鲁棒性提高。同时,该发明通过在模型中同时修正时序上下文依赖关系和多模态交互融合关系的方法,明显的提高了连续维度情感估计的准确度。



CN 113269277 B

1.基于Transformer编码器和多头多模态注意力的连续维度情感识别方法,包括如下步骤:

步骤一、对不同模态输入的特征,得到不同模态的嵌入特征表达;

将不同模态下提出的特征首先输入到不同的时序卷积网络中,得到不同模态下的短时特征表达,并利用正弦位置编码器对所述短时特征表达进行处理生成不同时刻的信息,与短时特征表达在时序上按位相加得到不同模态的嵌入特征表达;

步骤二、将不同模态的嵌入特征表达作为输入,利用多模态Transformer编码器得到不同模态的高级表征;

多模态Transformer编码器迭代使用三个子模块进行特征的学习,第一个是多模态多头注意力模块,第二个是时序多头注意力模块,第三个是前向传播模块,三个模块串联起来按顺序依次执行,最后将包含三个模块的多模态Transformer编码器迭代使用多次,其中多模态多头注意力模块可以对不同模态的特征进行动态的交互融合,其编码得到的不同模态的特征再输入时序多头注意力模块,获取时域中上下文的依赖,然后将编码了多模态和时序信息的特征输入前向传播模块进行非线性变化;通过对多模态Transformer编码器迭代使用,逐渐的修正多模态交互融合和时序上下文的依赖关系;

步骤三、将不同模态的高级特征表达作为输入,计算出每一时刻的情感状态值;

推理网络把多模态Transformer编码器输出的每一时刻的多模态特征拼接在一起,输入全连接层计出每一时刻的情感状态值。

基于Transformer编码器和多头多模态注意力的连续维度情感识别方法

技术领域

[0001] 本发明采用时序深度卷积神经网络(TCN)、自注意力Transformer编码器(Transformer Encoder)以及多模态多头注意力机制(Multi-modal Multi-head Attention),涉及一种从多模态(听觉、视觉)时序信息中对连续维度情感进行估计的模型和识别方法。

背景技术

[0002] 自动情感识别领域近年来越来越受到人们的关注,如在人机交互领域中,机器可以自动识别被观测者的情绪,并做出相应的反应。目前情感识别领域主要分为两类,一种是离散的情感识别,即将人的情感分类为高兴,悲伤,生气等等几种常见状态;另外一种是连续的情感识别,它将人的情感状态用两个维度进行表示,其中Arousal表示兴奋程度,Valence表示愉悦程度。正是因为连续情感可以更加精细描述人的情感状态,近年来对连续情感的识别成为了研究的热点。

[0003] 在过去几年中,通过音视频多模态来进行连续维度情感估计已经取得了许多重要的成果,并且大量的研究已经证明了基于多模态的连续情感识别方法效果要优于单模态的方法。文献“Multimodal Continuous Emotion Recognition with Data Augmentation Using Recurrent Neural Networks,2018th AVEC,pp57-64”公开了一种基于音频和视频的多模态连续维度情感估计方法。此方法使用经典的LSTM作为时序模型,得到时域上下文的依赖,并完成时间序列上的回归,得到每一时刻情感状态arousal/valence的估计。另外在多模态融合上,该方法使用了两种经典的融合方法,即特征融合和决策融合。但是,这种模型在连续维度情感估计阶段,由于LSTM模型在获取时域上下文依赖时对每一帧都进行了同样的处理,无法得到有重点的选则关键的上下文依赖信息,使得模型受到了一定的局限,导致对连续维度情感估计的准确率降低,泛化性能差,无法达到精度的要求;另外在多模态融合阶段,该方法受限与特征种类多,且无法动态实时的关注到重要模态的信息的限制,导致模型计算量大,且影响了模型的估计准确率,因此该方法具有一定的局限性且难以推广。

[0004] 目前研究学者已经在连续维度情感估计模型中取得了一定的成果,然而由于情感的复杂性和个体差异性,连续维度情感估计仍然面临以下挑战:

[0005] 1)“关键帧”问题。在长时序的连续维度情感估计任务中,每一时刻的情感状态与最近时刻的情感状态具有强相关性,且和某些关键时刻的情感信息具有更强的相关性,同时,每一时刻的情感状态和很久之前的情感信息可能关系较小。在过去的连续维度情感估计研究中,在对每一时刻的情感状态进行估计时,过去的情感信息都是以同等重要的方式进行处理,导致了模型难以获取关键的上下文信息,影响了模型的泛化能力和准确度。

[0006] 2)“多模态融合”问题。传统的多模态融合方法往往都局限于前期特征融合和后期决策融合两种方式,但是前期特征融合往往会导致特征维数高,容易过拟合,导致模型泛化能力差;对于后期决策融合,因为决策融合时的输入是不同特征回归后的结果,决策融合时

并未考虑不同特征之前的互补关系,因此决策融合方法往往难以挖掘不同模态之前的互补性。

[0007] 综上所述,现有的连续维度情感估计方法受到了时序模型的限制,难以发觉关键的时间上下文信息,另外在多模态融合时,大量多模态信息难以有效融合,容易造成连续维度情感估计精度低,泛化能力差等问题。

发明内容

[0008] 为了解决上面提到的这些问题,本发明设计了一种基于时序卷积神经网络(TCN),Transformer编码器(Transformer Encoder)、多模态多头注意力(MMA)的多模态连续维度情感估计模型及其识别方法,本发明的创新点如下:

[0009] 1)对长时序情感状态估计中时域上下文依赖,首先引入Transformer编码器的多头时序注意力模块来获取时域中信息的上下文依赖关系,为了排除时域上很久之前的信息带来的干扰,使模型更加关注最近一段时间中有效的上下文信息,本发明提出使用实时的掩码信息,作用于计算时序上的注意力关系,可以有效的解决时域上下文关系中的关键帧问题。

[0010] 2)提出了多模态多头注意力模块,在时序中每一时刻,每一个模态的信息都可以动态与其它模态进行交互,获取来自于其它模态中的互补的信息,完成了不同时刻不同模态重要性的判断和多模态信息的融合。提出的多模态多头注意力可以当作子模块与Transformer编码器中的时序注意力模态一起使用,将原来的Transformer编码器从时序上扩展到了多模态。

[0011] 3)提出一套完整的多模态连续维度情感估计模型,该模型共包含三个子网络:①特征嵌入网络,利用TCN提取不同模态的短时序特征表达,作为多模态时序编码器网络的输入;②多模态时序编码器网络,使用嵌入了多模态多头注意力的Transformer编码器,从输入的短时多模态特征,编码得到融合了时序上下文信息和多模态互补信息的高级特征表达;③推理网络,从多模态Transformer编码器输出的高级特征推理出当前情感状态。

[0012] 本发明解决其技术问题所采用的技术方案:时序卷积网络(TCN),Transformer编码器及和多头多模态注意力所组成的多模态连续维度情感识别模型,其特点如图1所示,该模型包括三个按先后顺序依次执行的子网络。具体的,本发明提出的基于Transformer编码器和多头多模态注意力的连续维度情感识别方法包括如下步骤:

[0013] 步骤一、对不同模态输入的特征,得到不同模态的嵌入特征表达(图1-Input Embedding Sub-network)。本发明中,将不同模态下提出的特征首先输入到不同的时序卷积网络中,得到不同模态下的短时特征表达,并利用正弦位置编码器(Sinusoidal Position Encoding)生成不同时刻的信息,与短时特征表达在时序上按位相加得到不同模态的嵌入特征表达。

[0014] 步骤二、将不同模态的嵌入特征表达作为输入,利用多模态Transformer编码器得到不同模态的高级表征(图1-Multi-modal Encoder Sub-network)。多模态Transformer编码器迭代使用三个子模块进行特征的学习,第一个是多模态多头注意力模块,第二个是时序多头注意力模块,第三个是前向传播模块,三个模块串联起来按顺序依次执行,最后将包含三个模块的多模态Transformer编码器迭代使用多次,其中多模态多头注意力模块可以

对不同模态的特征进行动态的交互融合，其编码得到的不同模态的特征再输入时序多头注意力模块，获取时域中上下文的依赖，然后将编码了多模态和时序信息的特征输入前向传播模块进行非线性变化。通过对多模态Transformer编码器迭代使用，可以逐渐的修正多模态交互融合和时序上下文的依赖关系。

[0015] 步骤三、将不同模态的高级特征表达作为输入，计算出每一时刻的情感状态值。推理网络把多模态Transformer编码器输出的每一时刻的多模态特征拼接在一起，输入全连接层(Fully Connect Layer)计出每一时刻的情感状态值(图1-Inference Sub-network)。

[0016] 本发明的有益效果是：通过使用Transformer编码器的时序多头注意力和实时的掩码信息对不同模态进行时域上下文信息的编码，可以更加的关注时序上过去某些关键时刻对当前情绪状态的影响，排除长远情感信息带来的干扰，使得模型鲁棒性提高。同时，该发明提出了一种多模态多头注意力模块，可以有效的嵌入到Transformer编码器中，从而挖掘了每一时刻下不同模态的关键信息。最后通过将时序注意力模块和多模态注意力模块联合迭代使用，可以逐步的修正不同模块的注意力信息，挖掘了有效的时序上下文和多模态信息。这种在模型中同时修正时序上下文依赖关系和多模态交互融合关系的方法，明显的提高了连续维度情感估计的准确度。

附图说明

[0017] 图1是本发明提出模型结构框图；

具体实施方式

[0018] 以下通过具体的实例对本发明的技术实施流程做进一步说明。

[0019] 1、不同模态的嵌入特征表达。

[0020] 本发明首先对每一模态的特征，使用1D卷积神经网络提取出30维的短时特征表达，然后使用位置编码器生成30维的不同位置的特征表达，然后将短时特征表达和位置特征表达进行按位相加，最后对每一模态得到30维的嵌入特征表达。

[0021] 2、多模态Transformer编码器提取高级表征。

[0022] 多模态Transformer编码器共包含3个顺序执行的子模块，下面对每一个子模块的实施做进一步说明。

[0023] a)多模态多头注意力模块(Multi-modal Multi-head Attention)

[0024] 多模态多头注意力模块主要用于获取多个模态之前的交互融合，如给定Q_j^t是模态j在t时刻下的特征向量，$Q^t = (Q_1^t, ..., Q_j^t, ..., Q_m^t)$是t时刻下所有模态组成的特征集合，因为多模态多头注意力基于自我注意力机制，因此我们定义$K_j = V_j = Q_j$，然后我们将Q_j，K_j，V_j利用线性投影到多个子空间中，并计算在每一个时刻下，不同模态特征之前的注意力权值，然后加权得到每一个模态下新的特征向量，最后所有子空间下的特征向量串联起来再次经过线性投影得到最后的特征表示。整个多模态多头注意力模块的计算公式如下：

[0025] $$Q_i^t = Concat(Q_1^t W_{1,i}^{(M)Q}, ..., Q_j^t W_{j,i}^{(M)Q}, ...Q_m^t W_{m,i}^{(M)Q})$$

[0026] $$K_i^t = Concat(K_1^t W_{1,i}^{(M)K}, ..., K_j^t W_{j,i}^{(M)K}, ...K_m^t W_{m,i}^{(M)K})$$

[0027] $$V_i^t = Concat(V_1^t W_{1,i}^{(M)V}, ..., V_j^t W_{j,i}^{(M)V}, ...V_m^t W_{m,i}^{(M)V})$$

[0028] $$A_i^t = Softmax(\frac{Q_i^t K_i^{tT}}{\sqrt{d_k}})$$

[0029] $$head_i^t = A_i^t V_i^t$$

[0030] $$MultiHead(Q^t, K^t, V^t) = Concat(head_1^t, ..., head_h^t)W^{(M)O}$$

[0031] b)时序多头注意力模块(Temporal Multi-head Attention)

[0032] 时序多头注意力模块主要用于获取单个模态下时序上下文的依赖，如给定模态$Modality_j$，我们定义Q_j^t是时刻t下模态j的特征向量，$Q_j = (Q_j^1, ...Q_j^t, ..., Q_j^n)$是整个视频序列的特征集合，因为时序注意力基于自我注意力机制，因此我们定义$K_j = V_j = Q_j$，然后我们将Q_j,K_j,V_j利用线性投影到多个子空间中，并在每一个子空间计算每一个模态中，时序上不同时刻的注意力权值，并加权得到每个子空间下每一时刻的特征向量，最后将所有子空间中的特征向量串联起来再次线性投影得到最后的特征表示。整个时序注意力模块的计算公式如下：

[0033] $$MultiHead(Q_j, K_j, V_j) = Concat(head_j^1, ..., head_j^h)W_j^{(T)O}$$

[0034] $$where\ head_j^i = Attention(Q_j^i, K_j^i, V_j^i)$$

[0035] $$and\ Q_j^i = Q_j W_{j,i}^{(T)Q}$$

[0036] $$K_j^i = K_j W_{j,i}^{(T)K}$$

[0037] $$V_j^i = V_j W_{j,i}^{(T)V}$$

[0038] 其中注意力(Attention)计算公式如下：

[0039] $$A = Softmax(\frac{QK^T}{\sqrt{d_k}})$$

[0040] Attention(Q,K,V)＝AV

[0041] 为了添加实时的掩码信息，将掩码矩阵M与上式计算出来的注意力矩阵A相乘，计算公式如下：

[0042] A＝MA

[0043] c)前向传播模块

[0044] 通过b)，我们得到了各模态下的特征表示，该特征表示融合了来自不同模态的信息及时间上下文信息，前向传播模块包括了两个线性映射和一个RELU非线性激活函数，其计算公式如下：

[0045] $FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2$

[0046] 其中x为不同模态输入的特征序列，该前向传播模块可以提高模型的非线性拟合能力，使得模型更高得到更好的准确率。

[0047] 3、前向推理网络估计情感状态

[0048] 由2中我们得到了不同模态下的特征表达，每一个模态下的特征表达融合了来自不同模态的信息以及时序上下文的信息，然后我们将不同模态的特征表征串联在一起，通过一个全连接层进行线性映射，进行最终的情感状态估计。

[0049] 本发明主要设计了多模态多头注意力模块,并将其插入到Transformer编码器中,将该编码器扩展为可以同时为多模态特征获取时序上下文依赖关系的模型,利用该模型,实现了一个连续维度情感估计的框架。经过对本发明在国际公开的连续维度情感识别数据库RECOLA(Remote Collaborative and Affective Interactions Database)上进行的实验检验,以对情感维度Arousal估计的CCC(Concordance Correlation Coefficient)值可以达到0.872,对情感维度Valence估计的CCC值可以达到0.714。

[0050] 具体算例如下,比如对视频音频两个模态,其输入的特征序列分别表示为$X_{video}=[X^1_{video},...,X^n_{video}]$和$X_{audio}=[X^1_{audio},...,X^n_{audio}]$,其中n表示特征序列的长度。多模态特征序列$X_{video}$,$X_{audio}$按顺序依次执行发明内容中的三个步骤,具体如下:

[0051] 步骤一,对不同模态的输入X_{video}和X_{audio},分别用不同的1D时序卷积神经网络计算,得到编码了短时特征表达的特征$X_{video,l}$和$X_{audio,l}$,然后利用正弦位置编码器生成位置向量P=[P^1,...P^n],与多模态的短时特征$X_{video,l}$和$X_{audio,l}$按位相加得到输出$X_{video,l,p}=[X^1_{video,l}+P^1,...,X^n_{video,l}+P^n]$和$X_{audio,l,p}=[X^1_{audio,l}+P^1,...,X^n_{audio,l}+P^n]$。

[0052] 步骤二,步骤二中包含迭代重复执行N次的三个按顺序执行的模块,在第一次迭代时,将步骤一的输出$X_{video,l,p}$和$X_{audio,l,p}$作为第一个子模块的输入,第一个子模块(多模态多头注意力)首先将其复制为Q_{video},K_{video},V_{video}和Q_{audio},K_{audio},V_{audio},然后重新组织生成新的Q,K,V,其中$Q^t=[Q^t_{video},Q^t_{audio}]$,$K^t=[K^t_{video},K^t_{audio}]$,$V^t=[V^t_{video},V^t_{audio}]$,并按公式MultiHead($Q^t$,$K^t$,$V^t$)对新组织的Q,K,V计算每一个时刻t下多模态特征之前的依赖关系,得到第一次迭代的输出$X_{v}'_{ideo}$和X'_{audio}并送入到第二个子模块(时序多头注意力),第二个子模块首先将其复制为Q_{video},K_{video},V_{video}和Q_{audio},K_{audio},V_{audio},然后按公式MultiHead(Q_j,K_j,V_j)计算每一个模态j在时序上的依赖关系,得到输出$X_{v}''_{ideo}$和X''_{audio},并送入到第三个子模块(前向传播模块),第三个模块对不同模态的输入,按照公式$FFN(X^{t,''}_j)$计算每个模态j中每一个时刻t下的特征,进行非线性变换,得到输出X''_{video}和X''_{audio}。第三个子模块的第一次迭代的输出X''_{video}和X''_{audio}然后作为输入重新输入到第一个子模块进行下次迭代,每一次迭代顺序运行三个子模块,共迭代N次,直到迭代结束得到步骤二的输出,我们将其表示为$E_{video}=[E^1_{video},...,E^n_{video}]$和$E_{audio}=[E^1_{audio},...,E^n_{audio}]$。

[0053] 步骤三将步骤二的输出在每一个时刻进行拼接,得到$E_{multimodal}=[E^1_{multimodal},...,E^n_{multimodal}]$,其中$E^t_{multimodal}=Concat(E^t_{video},E^t_{audio})$,最后对每一个时刻的$E^t_{multimodal}$利用一个全连接层计算得到最后的情感状态值。

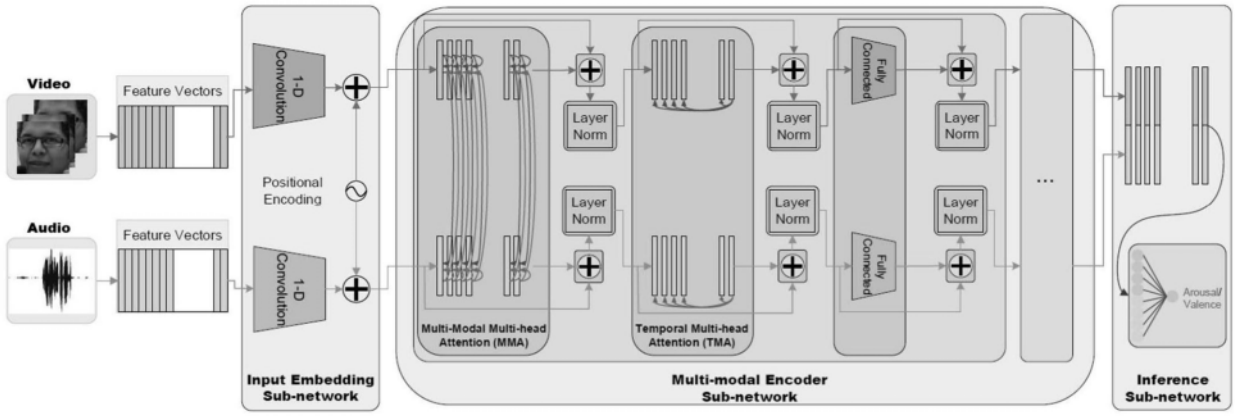

Video
Feature Vectors
Audio
Feature Vectors
1-D Convolution
Positional Encoding
1-D Convolution
Input Embedding Sub-network
Layer Norm
Layer Norm
Layer Norm
Layer Norm
Multi-Modal Multi-head Attention (MMA)
Temporal Multi-head Attention (TMA)
Fully Connected
Fully Connected
Layer Norm
Layer Norm
Layer Norm
Layer Norm
...
Multi-modal Encoder Sub-network
Arousal
Valence
Inference Sub-network

图1