季铭辉. 基于自注意力的时空关联性视觉情感分析模型[J]. 智能计算机与应用,2024,14(11):59-66. DOI:10.20169/j.issn.2095-2163.241108

基于自注意力的时空关联性视觉情感分析模型

季铭辉

(南京邮电大学 通信与信息工程学院, 南京 210003)

摘 要:针对视频数据中时间与空间维度特征难以融合的问题,本文提出一种基于多头自注意力与时空特征融合的视觉情感分析方法,使用改进型 Transformer 结构融合浅层特征图像序列中的深层时空间特征。首先,使用 CNN 网络提取图像帧序列中的浅层视觉特征;然后,使用多头自注意力机制提取深层空间特征,据此提取其时序特征,并使用残差结构融合时空间特征,以提取视觉模态中的深层情感特征信息;最后,使用分类网络预测视频样本的情感类别。实验结果表明,与传统的视频情感特征提取方法相比,该模型在识别视频情感方面能够获得更加优异的性能提升。

关键词:情感分析;时空特征融合;注意力机制;残差连接

中图分类号:TP183 **文献标志码**:A **文章编号**:2095-2163(2024)11-0059-08

Spatial-temporal attention-based visual sentiment analysis model

JI Minghui

(School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: To address the problem of the difficult fusion of temporal and spatial dimension features in video data, this paper proposes a visual sentiment analysis method based on multi-head self-attention and spatiotemporal feature fusion. An improved Transformer structure is used to fuse deep spatiotemporal features in shallow feature maps of image sequences. Firstly, a CNN network is utilized to extract shallow visual features from the image frame sequence. Then, a multi-head self-attention mechanism is employed to extract deep spatial features. Based on this, their temporal features are extracted, and the residual structure is used to fuse spatial and temporal features, enabling the extraction of deep emotion-related information in the visual modality. Finally, a classification network predicts the emotional category of the video sample. Experimental results demonstrate that compared to traditional methods for video emotion feature extraction, the proposed model can achieve superior performance improvements in recognizing video emotions.

Key words: sentiment analysis; spatial-temporal feature fusion; attention mechanism; residual connection

0 引 言

与文字和音频等其他载体相比,视觉信息能更直观地表现情感强度的变化,因此视觉情感分析成为近年来许多研究者的探讨方向。根据视觉数据形式的不同,通常可以将视觉情感分析分为图像情感分析与视频情感分析两个方面。相较于图像数据,视频数据具有丰富的时序消息,使得视频情感分析能够通过学习场景的细微变化捕捉到场景中事物在短时间内发生的情感变化,并且从动态的角度来展示情感信息,使得最终结果更加真实全面,也更具时效性。

传统机器学习算法需要将视觉信息转换为不同维度的视觉特征,包括几何特征、纹理特征、直方图信息等,使用情感模型将其映射至情感空间中,并通过模型的训练来实现视觉情感分析研究。主要有支持向量机、朴素贝叶斯、随机森林等算法[1-2]。然而此类方法需要巨大的标注数据集,且对生成的特征抽取较为依赖,导致其准确率和普适性有限,而随着卷积神经网络(Convolutional Neural Network, CNN)的迅速发展,越来越多的研究采用以 CNN 结构为基础的网络模型来对视觉情感进行分析处理,并取得了不错的研究成果[3]。与文本和音频情感分析可以使用循环神经网络处理序列信息不同,考虑到视

作者简介:季铭辉(1998—),男,硕士研究生,主要研究方向:人工智能。Email:1065095795@qq.com。

收稿日期:2023-05-22

频数据的复杂性,使用 CNN 结构处理视觉情感分析时通常需要事先将其切分成多个图像帧,而后再使用设计的 CNN 网络学习这些图像帧的情感表达。然而由于 CNN 网络本身仅考虑图像空间维度的局部特征变化,就使其无法很好地理解视频数据的时序信息。

近年来,注意力机制逐渐被应用在图像处理领域,现已在图像分类以及图像分割等任务中均取得了不错的效果,然而,要将注意力机制应用于视频处理方面还有诸多难题,亟待解决。

为解决上述问题,本文提出一种基于自注意力的时空关联性视觉情感分析模型(Spatial-Temporal attention - based Visual Sentiment Analysis Model, Spatial-Temporal Vision Transformer),旨在综合考虑视觉模态的时空间特征实现视觉情感类别的预测。该模型首先使用多头自注意力网络提取图像的空间特征,并针对多帧图像的空间特征,使用额外的多头自注意力网络学习其中的时序信息;接着,通过将各层网络输出的时间与空间特征相互拼接,得到最终的视觉情感特征;最后,在模型输出端连接上分类网络,完成数据样本的分类预测。在公开数据集上使用该模型与基线方法进行比较,用以验证其实际性能。

1 相关研究

CNN 本身具有的局部感知性和权值共享特点,使其能够有效地学习图像中的情感表示。文献[4]针对面部表情识别任务提出一种基于卷积神经网络以及软标签的面部表情识别框架,将多种情绪与每个表情相关联。但是,由于 CNN 仅考虑空间维度的局部特征相关性,而没有考虑时间维度的信息,导致其处理视频任务时面临一定的困难,所以通常需要使用额外的循环神经网络处理图像帧之间的时序信息[5-6]。除此以外,文献[7]在 CNN 的基础上设计出 3D-CNN 网络,使得卷积运算能够兼顾视频数据中的时序,从而更好地处理视频信息。文献[8]将 3D-CNN 网络应用于视频情感分析领域,使用 RNN 将 CNN 在单个视频帧上提取的外观特征作为输入,然后对运动进行编码,同时使用 C3D 网络对视频的外观和运动进行建模。文献[9]提出一种基于相关性的图卷积网络用于视频数据情感识别。该方法综合考虑类内和类间视频之间的相关性,并使用多头注意力机制来预测视频之间的隐藏关系,有效地提高分类的准确性。

与自然语言处理任务使用注意力机制相类似,在图像处理任务中,计算机同样需要关注图像中较为重要的部分,以利于更好地理解和处理图像中所包含的主要信息。文献[10]提出的 ViT(Vision Transformer)模型将 Transformer 模型引入到视觉领域,ViT 网络通过将图像像素矩阵分解成固定大小的图像块,然后将每个图像块展平成一维向量的形式,并通过添加位置编码与分类标签再将其组合成一个序列,使用 Transformer 模型中的 Encoder 模块进行处理,将分类标签对应位置上得到的输出向量作为分类任务的特征表示,最后通过全连接层映射到类别空间。与传统的卷积神经网络不同,ViT 在不使用任何卷积层的情况下,实现了与其相当的性能。

2 研究方法

本文提出一种基于自注意力的时空关联性视觉情感分析模型,其整体结构如图 1 所示。该模型首先接收连续 9 帧图像作为输入数据,使用 CNN 网络分别学习各帧图像的浅层空间特征表示,使用 Vision Transformer 网络分别提取其深层空间情感特征,并据此使用 Transformer 网络进一步抽取其中的时序特征,同时使用类残差结构将提取出的时间与空间特征进行融合,最终通过 Softmax 分类器预测情感极性。

2.1 空间情感特征表示

2.1.1 基于 CNN 的浅层特征提取

研究结果表明,时间维度上的连续多帧图像存在着时序依赖关系。既使将短视频样本处理成多帧图像数据,但由于每帧图像样本均需要输入到多头自注意力网络模型中进行训练,数据量依旧过于庞大且冗余,所以本文在输入端使用浅层卷积神经网络来提取各帧图像的浅层视觉特征,将高维度的图像数据映射到较低维度的特征空间,以简化表征图像的视觉信息,降低模型的计算复杂度。其卷积计算原理的数学公式为:

$$p(x,y)=\delta\cdot f(x,y)+b=\sum_{i=-a}^{a}\sum_{j=-b}^{b}\delta(i,j)f(x-i,y-j)+b \quad (1)$$

其中,$p(x,y)$ 表示经过卷积处理后的输出;δ 表示卷积核;$f(x,y)$ 表示待处理的原始输入;b 表示单个卷积核的偏移量。

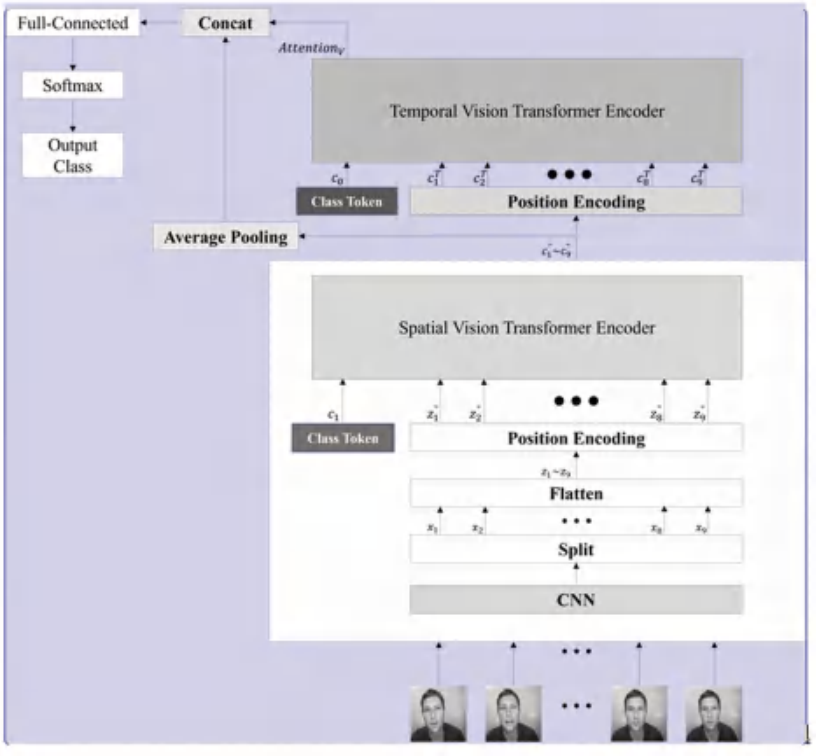

图 1　基于自注意力的时空关联性视觉情感分析模型

Fig. 1　Spatial-temporal self-attention-based visual sentiment analysis model

该卷积神经网络结构中，卷积核的步长 s 与填充 p 均为 1，每层卷积层后默认使用 ReLU 作为激活函数，池化层采用最大池化的方式，为了利于后续模型的处理，本文采用通道卷积的方式对输出特征的通道数量进行调整，通道卷积的卷积核大小为 1×1 的卷积，可以将其看作是对每个像素点或通道进行独立的线性变换，从而实现对特征通道的压缩或者扩展，经过通道卷积后，特征图像的通道数量变更为单通道。

2.1.2　基于多头自注意力的深层空间情感表示

Vision Transformer 网络提取图像空间特征需要将图像切分成固定大小的图像块，每帧图像经过浅层卷积神经网络抽取得到的图像特征共用该编码器模块，具体结构如图 2 所示。

通过对浅层卷积神经网络层输出的特征图像进行分割处理，将其切分成大小相等的 9 个特征图像块，并对其进行平滑处理，将其转换为固定维度的特征向量，计算方式如下：

$$\boldsymbol{x}_i^o = Flatten(\boldsymbol{feature}_{\mathrm{CNN}}) \tag{2}$$

其中，$\boldsymbol{feature}_{\mathrm{CNN}}$ 表示 CNN 输出端得到的浅层特征图像，$\boldsymbol{x}_i^o \in R^d$ 表示经过平滑处理后得到的特征向量。

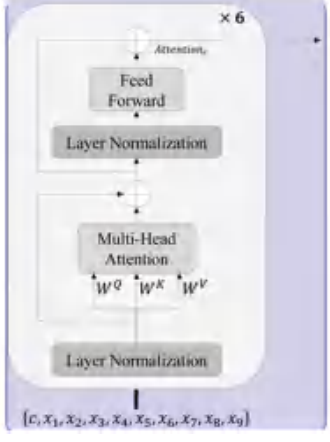

图 2　编码器结构

Fig. 2　Encoder structure

由于多头自注意力神经网络无法根据原始信息保证输入输出的位置匹配，因此需要提取原始图像块之间的相对位置信息，并将其嵌入特征向量序列之中。考虑到图像块存在空间关联性，故采用二维位置编码的方式，其位置信息的计算公式为：

$$PE_{(pos,t)} = \begin{cases} \sin\left(\dfrac{1}{10\,000^{4i/d}} \cdot pos\right), & t = 2i \\ \cos\left(\dfrac{1}{10\,000^{4i/d}} \cdot pos\right), & t = 2i + 1 \end{cases} \tag{3}$$

其中，$PE_{(pos,t)}$ 表示输入序列中 pos 对应的位置编码向量中位置 t 的元素，位置 t 为偶数时使用正弦函数计算编码，位置 t 为奇数时使用余弦函数计算编码；d 表示特征的维度；pos 表示向量所在的序列位置。分别计算 2 个维度的位置编码 $\boldsymbol{PE}_{pos_x}$，$\boldsymbol{PE}_{pos_y} \in R^{d/2}$，并将二者拼接得到最终需要嵌入的位置信息，计算公式如下：

$$\boldsymbol{PE}_{pos} = [\boldsymbol{PE}_{pos_x};\boldsymbol{PE}_{pos_y}] \in R^d \tag{4}$$

则多头自注意力网络的输入可以表示为：

$$\boldsymbol{X} = [\boldsymbol{c};\boldsymbol{x}_1;\boldsymbol{x}_2;\boldsymbol{x}_3;\boldsymbol{x}_4;\boldsymbol{x}_5;\boldsymbol{x}_6;\boldsymbol{x}_7;\boldsymbol{x}_8;\boldsymbol{x}_9] \tag{5}$$

$$\boldsymbol{x}_i = \boldsymbol{x}_i^o + \boldsymbol{PE}_{pos} \tag{6}$$

本文主要研究的是视觉情感分类任务，对于处理完成的特征向量序列，需要拼接一个与特征向量维度完全一致的分类标签 class token，用于分类任务中的类别预测。与其他特征向量序列有所不同，class token 属于序列输入中额外添加的一种特殊标记，不需要单独嵌入位置向量，因为其本质上已经被视为具有特殊位置编码的标记。

自注意力计算原理可由如下公式来描述：

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{W}^Q \boldsymbol{X} \\ \boldsymbol{K} = \boldsymbol{W}^K \boldsymbol{X} \\ \boldsymbol{V} = \boldsymbol{W}^V \boldsymbol{X} \end{cases} \tag{7}$$

$$\boldsymbol{Attention}_S = Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Softmax\left(\frac{\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{K}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{8}$$

其中，$\boldsymbol{W}^Q$、$\boldsymbol{W}^K$、$\boldsymbol{W}^V \in R^{t\times d}$ 分别表示计算 *Query* 向量、*Key* 向量以及 *Value* 向量的参数矩阵，$\boldsymbol{X}$ 表示嵌入位置编码并添加分类标签的输入序列。

多头自注意力机制计算公式可写为：

$$MultiHead(\boldsymbol{X}) = Concat(\boldsymbol{Attention}_{S1}, \boldsymbol{Attention}_{S2}, \cdots, \boldsymbol{Attention}_{Sn})\boldsymbol{W}^o \tag{9}$$

其中，$\{\boldsymbol{Attention}_{S1}, \boldsymbol{Attention}_{S2}, \cdots, \boldsymbol{Attention}_{Sn}\}$ 表示采用多组相互独立的参数矩阵 $\boldsymbol{W}^Q$、$\boldsymbol{W}^K$、$\boldsymbol{W}^V$ 计算得到的空间注意力输出矩阵；*Concat* 表示拼接操作，$\boldsymbol{W}^o \in R^{tn\times d}$ 表示映射矩阵。

2.2 基于多头自注意力的时序情感特征表示

针对 2.1 节输出的注意力矩阵，本节仅使用 class token 对应位置的输出作为时序信息获取的依据。由于每帧图像的分类标签特征向量之间只存在时序上的相关性，即向量之间的联系同文本内容里单词间的位置关联高度相似，即只存在横向间的联系，故而可以直接参照自然语言处理中 Transformer 模型的位置编码来进行操作，其位置编码方式如下：

$$PE_{(pos,t)} = \begin{cases} \sin\left(\dfrac{1}{10\ 000^{2i/d}} \cdot pos\right), & t = 2i \\ \cos\left(\dfrac{1}{10\ 000^{2i/d}} \cdot pos\right), & t = 2i+1 \end{cases} \tag{10}$$

其中，$PE_{(pos,t)}$ 表示输入序列中 pos 对应的位置编码向量中位置 t 的元素，位置 t 为偶数时使用正弦函数计算编码，位置 t 为奇数时使用余弦函数计算编码；d 表示特征的维度；pos 表示向量所在的序列位置。可以看出，该编码方案不需要后续的组合操作，计算得到的 $\boldsymbol{PE}_{pos} \in R^d$ 即位置信息编码。

同样，考虑到本实验的最终目的是情感分类，对嵌入位置信息的空间特征向量序列添加上分类标签，得到多头自注意力网络的输入 $[\boldsymbol{c}_0, \boldsymbol{c}_1^{\mathrm{T}}, \boldsymbol{c}_2^{\mathrm{T}}, \boldsymbol{c}_3^{\mathrm{T}}, \boldsymbol{c}_4^{\mathrm{T}}, \boldsymbol{c}_5^{\mathrm{T}}, \boldsymbol{c}_6^{\mathrm{T}}, \boldsymbol{c}_7^{\mathrm{T}}, \boldsymbol{c}_8^{\mathrm{T}}, \boldsymbol{c}_9^{\mathrm{T}}]$。为了方便后续模型的训练，需要对该分类标签进行初始化，其计算方式见如下：

$$\boldsymbol{c}_0 = \frac{\sum_{i=1}^{9} \boldsymbol{PE}_i}{9} \tag{11}$$

其中，向量 $\boldsymbol{PE}_i$ 的每个元素 $PE_{(i,t)}$ 由公式计算得到。

自注意力计算原理见下式：

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{W}^Q \boldsymbol{C} \\ \boldsymbol{K} = \boldsymbol{W}^K \boldsymbol{C} \\ \boldsymbol{V} = \boldsymbol{W}^V \boldsymbol{C} \end{cases} \tag{12}$$

$$\boldsymbol{Attention}_T = Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Softmax\left(\frac{\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{K}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{13}$$

其中，$\boldsymbol{W}^Q$、$\boldsymbol{W}^K$、$\boldsymbol{W}^V \in R^{h\times d}$ 分别表示计算 ***Query*** 向量、***Key*** 向量以及 ***Value*** 向量的参数矩阵，$\boldsymbol{C}$ 表示嵌入位置编码并添加分类标签的输入序列。

多头自注意力机制计算公式具体如下：

$$MultiHead(\boldsymbol{C}) = Concat(\boldsymbol{Attention}_{T1}, \boldsymbol{Attention}_{T2}, \cdots, \boldsymbol{Attention}_{Tm})\boldsymbol{W}^o \tag{14}$$

其中，$\{\boldsymbol{Attention}_{T1}, \boldsymbol{Attention}_{T2}, \cdots, \boldsymbol{Attention}_{Tm}\}$ 表示采用多组相互独立的参数矩阵 $\boldsymbol{W}^Q$、$\boldsymbol{W}^K$、$\boldsymbol{W}^V$ 计算得到的时间注意力输出矩阵；*Concat* 表示拼接操作；$\boldsymbol{W}^o \in R^{hm\times d}$ 表示映射矩阵。

2.3 时空间特征融合

考虑到模型学习空间特征序列的时间特征会不可避免地忽视部分关键的空间特征，因此本文将空间特征提取模块的输出序列经过平均池化后拼接在时间特征提取模块的输出端，构成一种类残差结构，

保留空间特征模块学习到的关键信息。平均池化公式如下：

$$\boldsymbol{c} = Pooling(\boldsymbol{C}) \tag{15}$$

残差结构计算方式具体如下：

$$\boldsymbol{feature} = Pooling(\boldsymbol{C}) + MultiHead(\boldsymbol{C})[class\ token] \tag{16}$$

2.4 Softmax 分类器

经由残差结构融合的时空间特征经过全连接层后，输入到 Softmax 分类器中，得到情感极性的概率分布。

全连接层的计算流程如下所示：

$$\boldsymbol{F} = FFN(\boldsymbol{feature}) = \boldsymbol{W}_2 \cdot \max(0, (\boldsymbol{W}_1 \cdot \boldsymbol{feature} + \boldsymbol{b}_1)) + \boldsymbol{b}_2 \tag{17}$$

其中，$\boldsymbol{W}_1$、$\boldsymbol{b}_1$ 和 $\boldsymbol{W}_2$、$\boldsymbol{b}_2$ 分别表示 2 层全连接层网络的权重矩阵与偏移向量。

Softmax 函数的基本原理是将原始向量转换为概率分布，通过比较各类别概率的最大值即可判别当前样本的最终分类。在神经网络的输出层中，通常使用 Softmax 分类器将网络输出的特征向量转化为表示分类概率的向量，然后根据分类概率来进行分类预测。具体来讲，经由全连接层输出的特征向量 $\boldsymbol{F}$，再经过 Softmax 分类器的输出如公式所示：

$$Y_k = \frac{e^{\boldsymbol{\delta}_k^{\mathrm{T}}\boldsymbol{F}}}{\sum_{g=1}^{G} e^{\boldsymbol{\delta}_g^{\mathrm{T}}\boldsymbol{F}}} \quad (k = 1, 2, \cdots, G) \tag{18}$$

其中，Softmax 分类器中存在可供训练的参数矩阵 $\boldsymbol{\delta}$；$\boldsymbol{\delta}_k$ 表示第 k 类所对应的分类器参数；G 表示类别数量，使用指数运算的形式是因为指数函数的导数始终为正值，数值较大的元素在经过指数函数的变换后差异将被放大，从而更加容易被分类器所区分。

2.5 损失函数

该模型使用交叉熵损失函数，假设模型输入为 X，输出类别为 Y，对于数据集中的每个样本 x_i 来讲，都有唯一的输出类别 y_{ture} 与之对应，使用 Softmax 分类器将样本的概率 x_i 输入模型得到的特征向量映射为概率分布，即目标 x_i 被判断为各类别的概率分布 $P(y = Y \mid x = i)$，具体公式如下：

$$P(y = Y \mid x = i) = \begin{bmatrix} p(y = 1 \mid x = i; \boldsymbol{\delta}) \\ p(y = 2 \mid x = i; \boldsymbol{\delta}) \\ \vdots \\ p(y = G \mid x = i; \boldsymbol{\delta}) \end{bmatrix} = \frac{1}{\sum_{g=1}^{G} e^{\boldsymbol{\delta}_g^{\mathrm{T}}\boldsymbol{F}_i}} \begin{bmatrix} e^{\boldsymbol{\delta}_1^{\mathrm{T}}\boldsymbol{F}_i} \\ e^{\boldsymbol{\delta}_2^{\mathrm{T}}\boldsymbol{F}_i} \\ \vdots \\ e^{\boldsymbol{\delta}_G^{\mathrm{T}}\boldsymbol{F}_i} \end{bmatrix} \tag{19}$$

模型训练的过程就是使用梯度下降的方式更新优化参数矩阵 $\boldsymbol{\delta}$，使得模型判定样本 x_i 为其真实分类 y_{ture} 的概率尽可能高，其损失函数的计算方式可表示为：

$$Loss(\boldsymbol{\delta}) = -\frac{1}{n}\left[\sum_{i=1}^{n}\sum_{j=1}^{G}(y_i \odot y_{\text{ture}}) \log \frac{e^{\boldsymbol{\delta}_j^{\mathrm{T}}\boldsymbol{F}_i}}{\sum_{g=1}^{G} e^{\boldsymbol{\delta}_g^{\mathrm{T}}\boldsymbol{F}_i}}\right] \tag{20}$$

其中，n 表示训练样本数量；“$\odot$”表示同或运算符，即如果模型预测类别 y_i 与样本真实类别 y_{ture} 一致，则 $y_i \odot y_{\text{ture}}$ 的计算结果为 1，反之则为 0。

3 实验与结果分析

3.1 数据集

本文使用 MELD[11]、CMU-MOSI[12]、CMU-MOSEI[13] 三个数据集对模型的性能进行验证实验，数据集中样本均被标注为积极、中性、消极三种情感极性类别。实验数据集详细信息见表 1。

表 1　数据集详细信息

Table 1　Dataset details

数据集	积极	中性	消极
MELD	2 400	2 400	2 400
CMU-MOSI	510	510	510
CMU-MOSEI	6 000	6 000	6 000

由于 CMU-MOSI、CMU-MOSEI 两个数据集具有相同的情感标注，范围从-3~3，其中-3 表示极端消极，-2 表示消极，-1 表示轻微消极，0 表示中性，+1 表示轻微积极，+2 表示积极，+3 表示极端积极。而 MELD 数据集的情绪标注只有积极、中性、消极，为统一数据标注，本文将情感标注在-3~-1 之间的样本作为消极样本，情感标注为 0 的样本作为中性样本，情感标注在 1~3 之间的样本作为积极样本。训练开始前，首先需要将数据集以 8:1:1 的比例划分为训练集、验证集和测试集，同时保证各样本集合中各情感分类样本的占比完全一致。

3.2 评价指标

评价指标用于衡量模型的性能与精度，在二分

类任务中，通过混淆矩阵计算出准确率(*Accuracy*，*A*)、精确率(*Precision*，*P*)、召回率(*Recall*，*R*)、综合评价分数(*F*1 - Score，*F*1)，可以全面地分析模型在分类任务中的性能表现。

二分类任务的混淆矩阵见表 2。表 2 中，*TP*(True Positive)表示被模型预测为正的正类样本；*FP*(False Positive)表示被模型预测为正的负类样本；*FN*(False Negative)表示被模型预测为负的正类样本；*TN*(True Negative)表示被模型预测为负的负类样本。各评价指标的计算公式具体如下：

$$\begin{cases} A = \dfrac{TP + TN}{TP + TN + FN + FP} \\ P = \dfrac{TP}{TP + FP} \\ R = \dfrac{TP}{TP + FN} \\ F1 = \dfrac{2 \times P \times R}{P + R} \end{cases} \tag{21}$$

表 2 二分类混淆矩阵

Table 2 Confusion matrix

真实标签	预测标签	
	正类样本	负类样本
正类样本	*TP*	*FN*
负类样本	*FP*	*TN*

准确率 *A* 表示被分类模型正确分类的样本数量与总样本数量的比率；精确率 *P* 表示被分类模型正确预测为正例的样本数量与分类模型预测为正例的样本数量的比率；召回率 *R* 表示被分类模型正确预测为正例的样本数量与实际正例的样本数量的比率；综合评价分数 *F*1 是精确率和召回率的调和平均值。

本文的研究内容是多分类任务，仅是简单地套用二分类评价指标并不能有效地分析模型性能，需要对所有的分类情况进行整体评估，所以本文使用宏平均(Marco Average)的方式处理原始评价指标，其基本原理是通过分别计算单一类别的评价指标，然后对各自得到的分数取平均值，以此作为模型综合性能的评判标准。

$$\begin{cases} P_{MA} = \dfrac{1}{k}\sum_{i=1}^{k} P_i \\ R_{MA} = \dfrac{1}{k}\sum_{i=1}^{k} R_i \\ F1_{MA} = \dfrac{1}{k}\sum_{i=1}^{k} F1_i \\ A = \dfrac{1}{SUM}\sum_{i=1}^{k} TP_i \end{cases} \tag{22}$$

其中，*k* 表示多分类任务中类别数目；TP_i 表示各分类样本中被正确判定的样本数量；*SUM* 表示测试数据样本的总数量；准确率 *A* 的计算方式虽然有所变化，但是其本质与二分类中准确率的计算原理并无区别，都是分析模型识别总体样本的正确概率。

3.3 实验设置

3.3.1 数据预处理

实验过程中需要对数据集中的短视频样本进行预处理。使用 OpenCV 库读取数据集中的视频文件，再将其切分成连续的图像帧，并将各帧图像按照时间顺序保存下来。从视频的第一帧开始，等间隔抽取其中 9 帧图像，该间隔根据图像帧的总数量决定，为使抽取的图像能够均匀分布在整段视频中，即如果总共有 26 帧，则会选取第 1、4、7、10、13、16、19、22、25 帧作为模型输入。使用 OpenCV 库将图像分辨率调整为 192×192。以上就是短视频样本的预处理过程。

3.3.2 实验参数设置

实验中，卷积神经网络模块总共包含 6 个卷积层，其卷积核大小除最后一层外均为 3×3，最后一层卷积层大小为 1×1，第 2 层与第 4 层卷积层后均采用最大池化，最终得到大小为 48×48 的特征图像，将特征图像切分成 9 个大小相等的图像块，并进行平滑处理，得到维度为 256、数量为 9 的特征向量。

空间多头自注意力模块使用 8 个 Multi-Head Attention 结构，即总共 8 组参数矩阵，每组参数矩阵 $\boldsymbol{W}^Q$、$\boldsymbol{W}^K$、$\boldsymbol{W}^V \in R^{32\times256}$，计算每个 Head 中的输出序列并将其转换为矩阵形式，即可得到尺寸均为 32×256 的 8 组子矩阵 $\{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3, \boldsymbol{Z}_4, \boldsymbol{Z}_5, \boldsymbol{Z}_6, \boldsymbol{Z}_7, \boldsymbol{Z}_8\}$，将 8 组子矩阵简单拼接得到 Multi-Head Attention 结构的最终输出矩阵 $\boldsymbol{Z} \in R^{256\times256}$。

时间多头自注意力模块使用 4 个 Multi-Head Attention 结构，即总共 4 组参数矩阵，每组参数矩阵 $\boldsymbol{W}^Q$、$\boldsymbol{W}^K$、$\boldsymbol{W}^V \in R^{64\times256}$，计算每个 Head 中的输出序列并将其转换为矩阵形式，即可得到尺寸均为 64×256 的 4 组子矩阵 $\{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3, \boldsymbol{Z}_4\}$，将 4 组子矩阵简单拼接得到 Multi-Head Attention 结构的最终输出矩阵 $\boldsymbol{Z} \in R^{256\times256}$。

分类模块中的全连接层后均需要添加 Dropout 结构，防止模型对于训练数据过度拟合，有效地提高模型的泛化能力。实验中将 Dropout 参数设置为 0.5，使用 Adam 优化器，设置初始学习率为 0.001，每经过 10 个 *epoch* 学习率会缩小 10 倍，模型的 *batch_size* 为 20，训练 *epoch* 设置为 100，每个 *epoch* 对

训练集进行随机打乱处理。

3.4　实验结果

为验证 Spatial-Temporal Vision Transformer 模型的有效性，实验中与一些基线模型在不同数据集上的性能进行比较，实验结果见表 3。

表 3　情感识别效果

Table 3　Results of sentiment recognition　　%

模型	MELD		CMU-MOSI		CMU-MOSEI	
	A	*F*1	*A*	*F*1	*A*	*F*1
C3D[14]	50.14	50.96	49.67	49.89	50.22	50.67
MMHA[15]	58.78	59.05	60.78	61.05	62.33	62.58
Graph-MFN[13]	64.32	65.40	67.32	67.40	68.38	68.56
Spatial-Temporal Vision Transformer	**65.56**	**65.57**	**69.28**	**69.34**	**70.39**	**70.42**
Spatial Vision Transformer	59.58	60.21	63.40	63.85	64.89	65.07
Temporal Vision Transformer	61.25	61.68	62.75	62.97	63.28	63.56

注：Spatial Vision Transformer 表示仅使用空间多头自注意力模块，Temporal Vision Transformer 表示仅使用时间多头自注意力模块。

从表 3 中可以看出，相比于传统 C3D 网络，Spatial-Temporal Vision Transformer 模型在不同数据集上的性能测试中，各项评价指标均有明显提升，主要原因在于 C3D 网络的整体结构较为简单，使其只能使用三维卷积的方式简单融合视频流中的时空间特征，导致其在提取情感特征的能力受到很大限制。

而 Spatial-Temporal Vision Transformer 模型相较于 MMHA 模型，准确率 *A* 平均提升约 8 个百分点，主要原因在于 MMHA 模型仅简单地将视频样本输入模型进行训练，模型中仅使用一种 Transformer 结构提取视频样本中的时空间特征，并没有针对性地考虑视频样本中的空间特征与时序特征的变化，使得注意力机制难以有效地分配视觉特征的权重，且由于限制输入图像帧的大小与数量，导致其无法学习到较为完整的视觉情感表示。而 Spatial-Temporal Vision Transformer 模型则使用 2 种不同的 Transformer 结构分别提取视频样本中的时空间特征，并将其拼接融合，既保证时空间特征提取的有效性，同时又因为结构之间的递进关系，增强了时空间特征的关联性，在测试过程中取得较大的性能提升。而 Spatial-Temporal Vision Transformer 模型相较于 MMHA 模型，综合评价分数 *F*1 提升幅度略微逊色于准确率 *A*，主要原因在于 MMHA 模型没有学习到完整的情感表示，导致针对 3 类情感类别的识别精确率波动幅度较大，而 Spatial-Temporal Vision Transformer 模型采用时空间特征融合的方式学习深层的视觉情感表示，在 3 类情感类别的识别精确率上较为稳定，导致其综合评价分数 *F*1 的提升也较为稳定。

而 Spatial-Temporal Vision Transformer 模型相较于 Graph-MFN 模型存在一定的性能提升，其准确率 *A* 平均提升约 2 个百分点，主要原因在于 Graph-MFN 模型通过引入注意力机制和可视化技术，可以更好地理解数据属性和特征简单融合模态特征，从而提高对异常数据和误差的鲁棒性，主要体现在针对本次研究输入图像帧的调整，Graph-MFN 模型性能没有受到太大影响。而由于 Graph-MFN 模型只是在图卷积网络中将视觉模态数据都视作标准的二维张量来处理，无法充分挖掘视频数据的时空间相关性，导致其在处理视频任务时，容易忽略部分重要的局部时空间特征，从而影响情感分类性能。而 Spatial-Temporal Vision Transformer 模型针对视频样本的特点抽取其中的视觉情感特征，故而性能优于 Graph-MFN 模型。而 Spatial-Temporal Vision Transformer 模型相较于 Graph-MFN 模型，综合评价分数 *F*1 提升幅度与准确率 *A* 大致相同，主要原因在于 Graph-MFN 模型虽然未能有针对性地考虑视觉模态的时空间特征变化，但由于其使用基于注意力机制的图卷积网络，在关键特征的提取方面明显优于 MMHA 模型，导致其针对 3 类情感类别的识别精确率波动幅度较小，综合评价分数 *F*1 的提升也较为稳定。但是从整体上来看，同样由于 Graph-MFN 模型未能有针对性地考虑视觉模态的时空间特征变化，导致其出现分析性能的瓶颈，使得其综合评价分数 *F*1 略微逊色于 Spatial-Temporal Vision Transformer 模型。

3.5　消融实验

为验证不同模块对整体模型的影响，实验还测试了各消融模型的性能。对比分析可知，相较于

Spatial Vision Transformer 模型, Temporal Vision Transformer 模型识别 MELD 数据样本时性能明显提升,而在识别 CMU-MOSI 和 CMU-MOSEI 数据样本时却出现性能略微损失的情况,主要原因在于 Temporal Vision Transformer 模型考虑空间特征之间时序关联性的同时,也间接性地丢失部分空间情感信息;而 MELD 作为多人物对话数据集,其分割的短视频依旧存在多个人物的情感变化,虽然 Spatial Vision Transformer 模型在训练过程中会尽可能地关注说话者的情感状态,但是其他人物的情感状态依旧会对模型分析造成影响。相反,使用 Temporal Vision Transformer 模型更能关注说话者在时序上的情感变化,所以模型分析性能提升明显。而在融合 Spatial Vision Transformer 模型与 Temporal Vision Transformer 模型得到的 Spatial-Temporal Vision Transformer 模型对比前二者在各个数据集上提升显著,说明 Spatial-Temporal Vision Transformer 模型能够有效地融合时间与空间维度的深层视觉特征,并表现出优异的测试性能。

4 结束语

本文从视频数据中时间与空间特征的关联性角度出发,设计出一种基于自注意力的时空关联性视觉情感分析模型,使用改进的 Transformer 网络结构融合浅层特征图像中的时间与空间特征,并且在时空特征融合模块中采用残差连接的方式进行特征融合,提取视觉模态的深层情感特征信息。实验结果证明,该模型能够更加有效地识别视觉数据的情感类别。

参考文献

[1] PIANA S, STAGLIANO A, ODONE F, et al. Real-time automatic emotion recognition from body gestures[J]. arXiv preprint arXiv,1402.5047, 2014.

[2] THOMAS R, RANGACHAR M J S. Fractional bat and multi-kernel-based spherical SVM for low resolution face recognition [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2017, 31(8): 1756014.

[3] ZHANG Tong, ZHENG Wenming, CUI Zhen, et al. A deep neural network-driven feature learning method for multi-view facial expression recognition [J]. IEEE Transactions on Multimedia, 2016, 18(12): 2528-2536.

[4] GAN Yanling, CHEN Jingying, XU Luhui. Facial expression recognition boosted by soft label with a diverse ensemble[J]. Pattern Recognition Letters, 2019, 125: 105-112.

[5] HU M, WANG H, WANG X, et al. Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks[J]. Journal of Visual Communication and Image Representation, 2019, 59: 176-185.

[6] EBRAHIMI K S, MICHALSKI V, KONDA K, et al. Recurrent neural networks for emotion recognition in video [C]// Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. Seattle,USA:ACM, 2015: 467-474.

[7] ÇIÇEK Ö, ABDULKADIR A, LIENKAMP S S, et al. 3D U-Net: learning dense volumetric segmentation from sparse annotation[C]// Proceedings of the 19th International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2016: 424-432.

[8] FAN Yin, LU Xiangju, LI Dian, et al. Video-based emotion recognition using CNN-RNN and C3D hybrid networks[C]// Proceedings of the 18th ACM International Conference on Multimodal Interaction. New York:ACM, 2016: 445-450.

[9] NIE Weizhi, REN Minjie, NIE Jie, et al. C-GCN: Correlation based graph convolutional network for audio-video emotion recognition[J]. IEEE Transactions on Multimedia, 2020, 23: 3793-3804.

[10] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv,2010,11929, 2020.

[11] PORIA S, HAZARIKA D, MAJUMDER N, et al. Meld: A multimodal multi-party dataset for emotion recognition in conversations[J]. arXiv preprint arXiv,1810.02508, 2018.

[12] ZADEH A, ZELLERS R, PINCUS E, et al. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos[J]. arXiv preprint arXiv,1606.06259, 2016.

[13] ZADEH A, LIANG P, PORIA S, et al. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia:ACL, 2018: 2236-2246.

[14] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks [C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ:IEEE,2015: 4489-4497.

[15] CHEN Xi, LU Guanming, YAN Jingjie. Multimodal sentiment analysis based on multi-head attention mechanism [C]// Proceedings of the 4th International Conference on Machine Learning and Soft Computing. New York:ACM,2020: 34-39.