控制与决策
Control and Decision

基于混合注意力机制的多信息行人过街意图预测

桑海峰, 刘玉龙, 刘泉恺

引用本文:
桑海峰, 刘玉龙, 刘泉恺. 基于混合注意力机制的多信息行人过街意图预测[J]. 控制与决策, 2024, 39(12): 3946-3954.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2023.1406

基于混合注意力机制的多信息行人过街意图预测

桑海峰†, 刘玉龙, 刘泉恺

(沈阳工业大学 信息科学与工程学院，沈阳 110870)

摘 要: 提前预测道路两旁的行人是否存在过街意图或一段时间后是否会出现过街行为是自动驾驶汽车面临的重要挑战之一,如何有效融合不同模态的多元信息是准确预测行人过街意图的重要问题. 基于此,提出一种基于混合注意力机制的多信息融合预测模型,使用一种基于交叉注意力机制的图像特征融合网络来提取原始图像与语义图像之间的互补信息,并使模型更加关注与行人过街行为有关的图像部分. 同时,提出一种融合注意力机制的分级GRU模块,用以捕捉不同模态的非视觉信息对行人过街意图的影响. 在PIE和JAAD数据集上对所提模型进行对比实验,已验证其具有领先于同类研究的性能;针对所提出模块进行广泛的消融实验,表明其有效性.

关键词: 行人过街意图预测；交叉注意力机制；自动驾驶；视频分析；计算机视觉；多信息融合

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2023.1406

引用格式: 桑海峰, 刘玉龙, 刘泉恺. 基于混合注意力机制的多信息行人过街意图预测 [J]. 控制与决策, 2024, 39(12): 3946-3954.

Multi information pedestrian crossing intention prediction based on mixed attention mechanism

SANG Hai-feng†, LIU Yu-long, LIU Quan-kai

(School of Information Science and Engineering，Shenyang University of Technology，Shenyang 110870，China)

Abstract: Predicting in advance whether pedestrians on both sides of the road have the intention to cross the street or whether crossing behavior will occur after a period of time is one of the important challenges facing self-driving cars. How to effectively fuse the multi-information from these different modalities is an important issue in accurately predicting pedestrian crossing intentions. Therefore, this paper proposes a multi-information fusion prediction model based on a hybrid attention mechanism. The model uses an image feature fusion network based on a cross-attention mechanism to extract complementary information between the original image and the semantic image and to make the model more attentive to the parts of the image that are relevant to the behavior of the pedestrian crossing the street. We also propose a hierarchical gated recurrent unit (GRU) module incorporating an attentional mechanism to capture the effects of different modalities of non-visual information on pedestrian crossing intentions. Finally, the proposed model is compared on the PIE and JAAD datasets and achieves leading performance, and extensive ablation experiments are conducted on the proposed module to prove its effectiveness.

Keywords: prediction of pedestrian crossing intention；cross attention mechanism；autonomous driving；video analysis；computer vision；multi-information fusion

0 引 言

自动驾驶作为汽车与人工智能领域的重要发展方向之一,近年来受到了众多高校及企业研究人员的关注. 行人作为交通参与者中最脆弱的一环,其安全理应受到其他所有交通参与者包括自动驾驶汽车的着重保护,自动驾驶汽车要在检测到行人位置与运动轨迹的基础上进一步关注于行人未来的运动趋势. 行人的过街行为是自动驾驶汽车与行人最容易发生冲突的关键环节,因此,为了尽最大可能保护道路两旁行人的安全,对行人过街意图的预测是高级别自动驾驶汽车的必备能力之一.

对于一个有过街意图的行人而言,其过街行为的发生与否受到多方面信息的影响,自动驾驶车辆除了分析行人的历史轨迹,对环境信息的融入也是必要的. 此外,使用行人骨骼关键点数据来提取行人运动特征也是预测行人未来动作的有效方式. 如何

收稿日期: 2023-10-08；**录用日期:** 2024-03-05.

基金项目: 国家自然科学基金项目(62173078)；辽宁省自然科学基金项目(2022-MS-268).

†通讯作者. E-mail: sanghaif@163.com.

融合影响行人过街的各个因素信息进行行人过街意图的预测,是近年来行人过街意图预测的重要研究方向. 文献[1]通过多个门控循环单元(gated recurrent unit, GRU)[2]模块按特征复杂程度分级融合图像信息、骨骼关键点信息、历史轨迹信息和车速信息;文献[3]提出的基准模型使用多个RNN (recurrent neural network)网络提取历史轨迹、骨骼关键点和车速信息,再与使用3D卷积神经网络(3D convolutional neural networks, 3DCNN)[4]提取出的行人局部视频序列特征相融合;文献[5-6]将所有信息分为视觉信息(原始图像和语义图像)和非视觉信息(历史轨迹、骨骼点和车速)分别融合,然后再进行视觉信息与非视觉特征的融合,取得了先进的预测结果. 然而,这些模型在提取视觉信息(原始图像和语义图像)过程中,只是分别提取出相关特征然后进行融合,并没有详细考虑原始图像信息与语义图像信息之间的互补性和冗余性. 此外,现有对非视觉信息的特征提取方法并不能使模型充分学习到各非视觉信息对行人过街意图的综合影响.

本文针对视觉信息中的原始图像信息和语义信息之间的交互作用进行研究,探索如何更好地将语义信息映射到原始图像信息中,并帮助模型专注于一些对行人过街行为产生影响的图像部分. 同时,本文提出一种融合注意力的分级GRU模块进行非视觉信息的特征提取. 本文主要贡献有:

1) 结合现有基准模型和交叉注意力机制提出一种新的混合注意力机制行人过街意图预测模型.

2) 在提取视觉信息特征时设计一种针对同一场景下的多信息图像进行特征提取与信息交互的交叉融合网络,可以有效剔除多信息图像中的冗余信息,并保留其互补信息,从而对该场景进行更加精确和丰富的特征提取.

3) 在提取非视觉信息特征时提出一种融合注意力机制的分级GRU模块,在有效编码各序列信息的基础上,准确捕捉到几种非视觉信息对于行人过街行为的互补作用.

1　相关工作

1.1　多信息行人过街意图预测

基于驾驶视角的行人过街意图预测最早出现在文献[7]中, Schneemann等提出了一种基于周围环境与行人运动的支持向量机(support vector machine, SVM)分类方法来预测行人过街意图. Rasouli等在文献[8]中发布了自动驾驶联合注意力数据集(JAAD),并使用卷积神经(convolutional neural networks, CNN)网络为基线就各种背景环境因素对行人过街行为的影响进行了研究. 但文献[8]只考虑了单独一帧的环境信息,忽略了环境和行人的运动特征. 随着循环神经网络的发展与成熟,文献[9-10]均使用了循环网络对行人过街行为进行时序建模;文献[11-12]将行人骨骼关键点引入到行人过街意图预测问题中,行人骨骼关键点序列可以很好地反应行人姿态及运动方式,通过从裁剪得到的行人图像中提取骨骼关键点位置,再使用手工特征或图卷积(graph convolutional networks, GCN)的方式可以提取到更为准确的行人运动特征,大幅提高行人过街意图预测的精度.

对于视频分析类任务,空间特征和时序特征都是影响结果的重要因素,文献[13-14]使用CNN与RNN相结合的方式分别提取空间特征和时序特征,文献[13]所提出的PIE数据集在之后的工作中获得了广泛的应用;文献[15-16]均直接采用3DCNN提取视觉图像序列信息进行时空建模预测行人过街意图.

语义信息作为理解图像内容的重要因素也被引入到分析行人过街意图上,文献[17-19]等都在使用原始视频帧图像的基础上,提取语义图像来帮助模型理解图像中各元素语义特征关系;文献[20]又引入光流信息与其他非视觉信息一起进行编码.

1.2　多信息融合方法

最近的相关研究工作大多将重点集中在如何融合视觉特征(原始图像、语义图像)、非视觉特征(行人检测框序列、行人骨骼关键点序列、车速)等多元信息上. 如文献[1]首先使用CNN提取行人外观特征和行人周围局部视觉特征,再通过5个GRU模块分级融合编码图像特征、骨骼关键点序列、检测框序列和车速序列;文献[21]使用长短期记忆(long short-term memory, LSTM)对各项信息进行序列编码后使用分级的全连接网络融合特征.

注意力机制作为机器学习模型中一种特殊结构,它以能够聚焦重要信息、忽略无关信息的重大优势在计算机视觉领域受到了广泛的应用,其应用在图像特征提取上可以使模型关注到图像中与任务相关的部分,应用在序列信息上可以知道事件发生的关键时间节点. 文献[22]使用卷积块注意力模块(convolutional block attention module, CBAM)提取视觉特征,使用GRU提取非视觉特征,使用注意力机制融合视觉与非视觉特征. 文献[3]提出了一个被广泛引用的基线模型,使用3D卷积提取视觉特征,使用

RNN和注意力机制提取非视觉信息,最后再用注意力机制融合. 文献[3]还为行人过街意图预测任务规定了统一的评估标准,使现有模型算法可以在该标准下进行对比. 文献[19]使用类似文献[1]的分级GRU的方式提取非视觉信息特征,使用CNN与GRU两步提取视觉信息的空间特征和时序特征,再将视觉特征和非视觉特征分别使用时间注意力机制进行降维后用注意力机制融合. 文献[23]在文献[19]的基础上添加交通信号灯信息与非视觉信息一起编码.

Transformer[24]架构在语言翻译和其他自然语言处理任务方面有重大改进,文献[25]中尝试使用了RubiksNet[26]和基于Transformer架构的Timesformer模型进行视频信息的编码,并使用Transformer编码非视觉信息,最后再使用注意力机制进行融合.

由上述文献可知,目前对于行人过街意图预测的研究所用信息来源很多,其中视觉信息包括原始视频图像帧、经过语义分割得到的语义视频图像帧,非视觉信息包括提取到的行人骨骼关键点序列、行人检测框序列、车速序列. 现有工作只是将这些特征分为视觉特征和非视觉特征分别进行编码,并未考虑视觉特征与非视觉特征内部信息的差异性和互补性. 例如,文献[17-19]中分别提取或裁剪出原始图像特征、语义图像特征、行人周边环境图像特征等,再将提取出的特征向量进行拼接融合;文献[3,22]则使用多分支的RNN分别编码行人骨骼关键点序列、行人检测框序列、车速序列,最后拼接融合.

针对现有方法无法提取到视觉特征与非视觉特征内部之间的信息互补与差异问题,本文引入交叉注意力机制进行研究. 交叉注意力机制已经在多模态学习、图像描述等领域取得了显著成果,交叉注意力机制可以在多个不同来源或不同模态信息之间进行信息交互,从而获得更加丰富的信息表示. 此外,本文沿用文献[3]中所用时间注意力机制进行时间维度的特征融合,再结合本文设计的基于交叉注意力机制的视觉特征融合网络、融合注意力机制的分级GRU模块,混合使用这些不同种类的注意力实现方式和使用方法[27],提出一种基于混合注意力机制的多信息行人过街意图预测模型.

2 基于混合注意力机制的多信息行人过街意图预测

行人过街意图预测的任务为:通过车载相机与传感器,观测行人0.5 s(约16帧),预测1.0 s后是否会出现过街行为. 本文提出了基于混合注意力机制的多信息行人过街意图预测模型,主要由视觉特征提取模块、视觉特征融合模块、非视觉特征融合模块、视觉与非视觉融合模块组成.

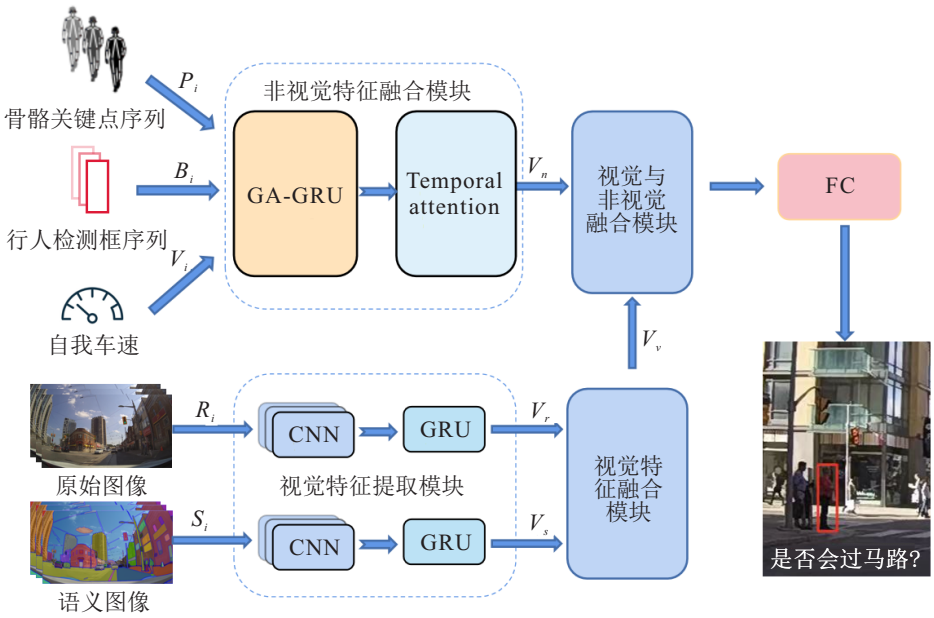

图 1 基于混合注意力机制的多信息行人过街意图预测模型

模型结构如图1所示,原始图像信息和语义图像信息会分别经过视觉特征提取模块中的CNN网络和GRU提取每帧图像的空间特征和视频序列的时序特征,再由基于交叉注意力机制的视觉特征融合模块将原始图像信息和语义图像信息相融合. 由于原始图像和语义图像在图像大小、图像内容上存在较强的相关性,交叉注意力机制可以很好地融合二者的互补信息,使模型对视觉图像内容理解更加丰富. 在对非

视觉信息(行人骨骼关键点、行人检测框、车速)进行特征编码时,本文借鉴了文献[1,18]中的分级GRU结构. 该结构首先对较为复杂的骨骼关键点信息进行编码,再逐步融合行人检测框信息和车速信息,本文在此基础上对各级特征融合结果进行注意力机制的计算,使3种不同模态的信息能互相关联融合,并减弱了行人骨骼关键点检测不全的影响.以下是对各模块的详细介绍.

2.1　模型输入信息

模型所需输入信息包括原始图像信息、语义图像信息、行人骨骼关键点信息、行人检测框信息、车速信息几个部分.

2.1.1　原始图像信息

在每个序列样本中,本文对每帧原始图像(大小为[1 920, 1 080, 3])进行了截取,以减少背景无关信息(天空、建筑、树木等)对模型的干扰,截取方法为取所观测目标行人检测框中心点所在的四分之一图像作为模型原始图像输入. 在该部分图像中,会存在充足的目标行人外观信息及行人周围环境信息. 截取后每张视频帧图像大小为[960, 540, 3]. 又由于本文所用CNN为预训练好的VGG16模型,需要原始图像首先改变宽高均为224. 使用i表示每段视频样本的编号,R_i表示截取处理后每段视频帧中的原始图像信息. R_i的数学表示为

$$R_i = \{r_i^{t-m}, r_i^{t-m+1}, \ldots, r_i^t\}. \tag{1}$$

其中:t为观测样本序列的最后一帧;m为观测样本序列的帧数;r_i为样本序列中每帧图像经过裁剪和归一化宽高后的RGB图像,其维度大小为$[m, 224, 224, 3]$.

2.1.2　语义图像信息

本文将每个样本中所有帧图像通过Deeplabv3模型提取到样本语义图像,Deeplabv3模型输入信息为[1 920, 1 080, 3]的原始图像,输出同等大小的彩色语义图像,对语义图像做与原始图像相同的裁剪,最后为了将行人走位环境语义凸显出来,本文对行人检测框内部分的图像进行掩膜处理,将其像素值设为白色. 最终,在经过图像宽高的归一化后,输入模型的语义图像大小同样为$[m, 224, 224, 3]$. 使用S_i表示语义图像信息,S_i的数学表达式为

$$S_i = \{s_i^{t-m}, s_i^{t-m+1}, \ldots, s_i^t\}. \tag{2}$$

2.1.3　行人骨骼关键点信息

人体骨骼关键点信息被广泛应用于人体动作识别与动作预测领域,对行人的骨骼关键点序列信息进行观测能有效预测行人的动作趋势. 本文模型所用的行人骨骼关键点信息是使用OpenPose预先提取出的行人18个关键点在行人检测框内的相对位置$P\,[x_1, y_1, x_2, y_2, \ldots, x_{18}, y_{18}]$. 每段视频帧样本的骨骼关键点信息维度为$[m, 36]$. 使用$P_i$表示骨骼关键点信息,$P_i$的数学表达式为

$$P_i = \{p_i^{t-m}, p_i^{t-m+1}, \ldots, p_i^t\}. \tag{3}$$

2.1.4　行人检测框信息

输入模型的行人检测框信息为行人检测框左上角点与右下角点的像素坐标$[x_1, y_1, x_2, y_2]$,每段视频帧样本的行人检测框信息维度为$[m, 4]$. 使用B_i表示语义图像信息,B_i的数学表达式为

$$B_i = \{b_i^{t-m}, b_i^{t-m+1}, \ldots, b_i^t\}. \tag{4}$$

2.1.5　车速信息

输入模型的车速信息来源为数据集内标注的车辆车速传感器记录的车速信息,单位为km/h,每帧车速信息维度为$[m, 1]$. 使用V_i表示语义图像信息,V_i的数学表达式为

$$V_i = \{v_i^{t-m}, v_i^{t-m+1}, \ldots, v_i^t\}. \tag{5}$$

2.2　视觉特征提取模块

本文使用在ImageNet数据集上预训练的VGG16模型和一个GRU来作为视觉特征提取器,每段观测到的视频帧序列作为一个4D矩阵输入视觉特征提取器,其输入尺寸为$[m, 224, 224, 3]$. 本工作中观测帧数m取为16,VGG16对输入图像进行特征提取后,使用一个大池化层将每张图像池化为一个$[1, 512]$大小的特征向量,对一个连续观测片段16帧的视频样本进行特征提取即得到一个$[16, 512]$的视频样本特征. 将得到的视频样本特征通过一个包含256个隐藏单元的GRU,得到最终的视觉特征V_r,其大小为$[16, 256]$. 对语义图像进行相同的视觉特征提取,得到同为$[16, 256]$大小的语义图像特征V_s. 该过程的公式表达为

$$V_r = \mathrm{GRU}_{nh=256}(\mathrm{VGG}(R_i)), \tag{6}$$

$$V_s = \mathrm{GRU}_{nh=256}(\mathrm{VGG}(S_i)). \tag{7}$$

2.3　视觉特征融合模块

视觉特征融合编码的目的是将原始图像序列特征和语义图像特征进行融合,实现原始图像与语义图像之间的信息互补,并使模型能够更多地关注到与影响行人过街意图有关的视觉特征.

现有工作多通过对不同视觉特征进行拼接然后使用注意力机制进行融合,但这种先拼接后融合的方式,忽略了多种视觉特征内存在的信息冗余与互补. 为了缓解这一问题,引入交叉注意力机制进行研究. 交叉注意力机制是一种很好的融合不同特征的方法,相比于传统的先拼接后融合的方式,交叉注意力可以在融合的过程中进行注意力机制的运算,两个不同特征相乘的方式可以更加直接有效地进行信息的交互共享,并赋予融合特征以权重,找到融合特征中与本文任务相关的重要特征. 交叉注意力机制实现方法众多,本文探索了几种不同交叉注意力机制的实现方式.

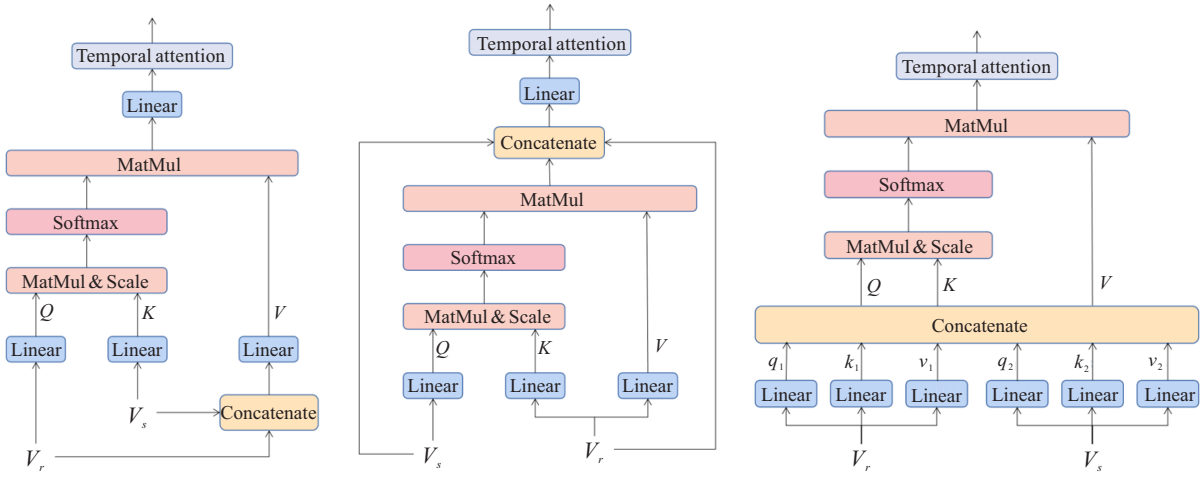

图 2 3种交叉注意力实现方式

如图2(a)所示,对视觉特征提取模块所提取到的视觉特征V_r、V_s分别做线性变换得到注意力机制所需的Q和K,同时将V_r与V_s相拼接进行线性变化得到注意力机制的V. 图2(b)则是对V_s进行线性变化得到Q,对V_r做两次不同的线性变化得到K与V,此外图2(b)还将原始的V_r和V_s与注意力机制所得结果相拼接以达到保留部分原始特征的目的.

如图2(c)所示,将两特征分别做3个线性变换,再将所得q_1、q_2拼接得到Q,同理得到K与V,然后进行注意力机制,最后使用一个线性层对融合特征的维度进行处理.

$$Q = \text{Concatenate}(Q_r, Q_s), \tag{8}$$

$$K = \text{Concatenate}(K_r, K_s), \tag{9}$$

$$V = \text{Concatenate}(V_r, V_s), \tag{10}$$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right). \tag{11}$$

上述3种交叉注意力实现方式最终都使用文献[2]中时间注意力进行时序特征提取.

2.4 非视觉特征融合模块

文献[18]中使用的一种分级GRU的方式对模型输入的非视觉信息进行分级编码获得了不错的效果,但连续3个GRU的使用,容易造成第1个GRU输入信息的遗忘,因此设计如图3所示的融合交叉注意力机制的GRU(cross-attention GRU, CA-GRU)作为非视觉信息融合编码器,它能在混合编码3种非视觉信息的同时,在不同混合层次间进行交叉注意而不是只关注于最终的融合结果. 本文将三级GRU模块提取到的视觉特征分别作为注意力机制的Q、K、V输入进行注意力,所用GRU模块隐藏单元个数均为256,最后使用文献[3]中时间注意力机制进行时序特征的提取.

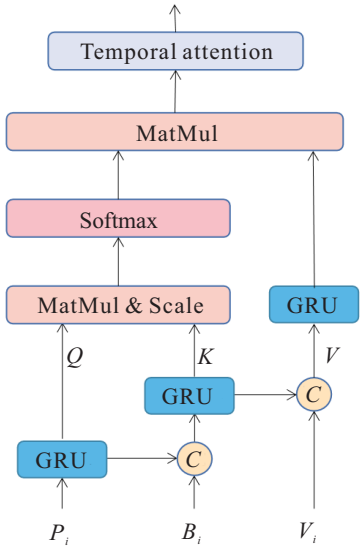

图 3 融合注意力机制的分级GRU模块

2.5　视觉与非视觉融合模块

本文使用一种以交叉注意力为基础的模态注意力机制融合视觉特征与非视觉特征, 由文献[3]中所用时间注意力变换而来, 其整体结构如图4所示. V_v 与 V_n 分别为视觉特征与非视觉特征, 其维度大小均为[1, 256], 该模态注意力机制同时对视觉特征和非视觉特征与拼接起来的总特征(总特征维度为[2, 256])进行注意力机制, 而不是单独对某一特征或融合特征进行注意力, 以得到视觉特征与非视觉特征同时作用于行人时的相关权重. 最后拼接并使用线性层降维, 视觉与非视觉融合模块输出结果维度为[1, 256].

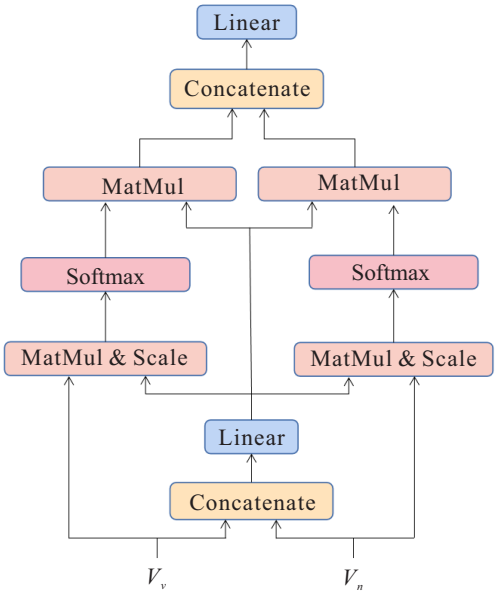

图 4　视觉特征与非视觉特征融合模块

2.6　损失函数

本文将行人过街意图看作一个二分类问题, 模型输出为一个概率值, 表示一段时间后行人发生过街行为的概率. 故训练过程中选用二元交叉熵损失函数(binary cross entropy), 其计算公式为

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1-y_i)\cdot\log(1-p(y_i)). \tag{12}$$

其中: N 为样本总数, y 为每个样本的真实标签(取值0或1), $p(y_i)$ 为每个样本预测为1的概率.

3　实　验

由于街道场景的复杂性和行人运动的不确定性, 自动驾驶汽车想要提前较长时间准确预测行人的过街行为几乎是不可能的, 但在得出预测结果后, 给自动驾驶汽车预留足够的制动或减速时间是必要的. 目前多数工作的做法是在行人过街行为开始的前1～2 s时进行预测, 即行人过街意图预测模型所观测的视频序列的最后一帧在行人开始过街前1～2 s. 以行人开始过街时刻作为第0帧原始帧, −30帧到0帧为预留制动时间, −45帧到−30帧为模型观测的视频帧序列.

3.1　数据集与评价指标

PIE(行人意图估计)数据集是自动驾驶领域最常用的数据集之一, 它包含超过6小时的驾驶视角视频信息, 对视频中车辆、行人、交通信号灯等边界框进行了标注, 还包括行人意图以及车辆传感器信息. 在PIE数据集长度合适的每个行人视频序列中采集6个样本, 相邻样本间重复率为0.6, 共采集4 770和3 816个样本用作训练和测试.

JAAD(自动驾驶联合注意力)数据集是一个驾驶视角下的行人数据集, 由346个短视频片段组成, 它同样包含检测框和动作信息的标注, 但JAAD数据集缺少车辆的自我运动信息. 本文继承现有工作的试验方法, 使用JAAD数据集的两个变体, JAAD_{beh} 和 JAAD_{all}, 其中 JAAD_{beh} 共2 134个训练样本与1 881个测试样本, JAAD_{all} 有8 613个训练样本与6 732个测试样本.

为了准确衡量模型性能, 本文使用ACC准确率、曲线下面积AUC、Precision、F1、Recall这些性能指标, 并与现有工作进行对比.

3.2　实验设置

本文所用GRU以及模态注意力隐藏元个数均为256, 并在最终的预测层使用0.001的L2正则化, 车速信息只在PIE数据集上使用, 在PIE数据集上训练时学习率为0.000 005, 共训练80个epoch. 在JAAD数据集上学习率设为0.000 05, 共训练150个epoch. 本文所有实验基于Ubuntu20.04LTS系统, 所用硬件设备为NVIDIA GeForce RTX 3090显卡, 设备内存32 GB.

本文对比的基线模型均为经典的PIE过街意图预测模型, 包括使用多个分级GRU融合不同信息的SF-GRU[1]、使用多个RNN和注意力分别处理不同信息的基准模型PCPA[3]、文献[18]中提出的PCIP模型, 应用Transformer[24]进行特征融合的CAPformer[25], 以及引入光流信息的STFF-MANet[20]模型.

3.3　实验结果与分析

3.3.1　实验结果

表1为本文模型与现有工作的实验对比结果. 可

表 1 本文模型与基线模型结果相比

models	PIE					$JAAD_{beh}$					$JAAD_{all}$				
	ACC	AUC	Prec	F1	Recall	ACC	AUC	Prec	F1	Recall	ACC	AUC	Prec	F1	Recall
SF-GRU[1]	0.83	0.78	0.50	0.63	0.78	0.53	0.54	0.58	0.65	0.62	0.76	0.77	0.40	0.53	0.79
PCPA[2]	0.85	0.86	0.69	0.77	**0.88**	0.55	0.52	0.63	0.65	0.67	0.76	0.79	0.41	0.55	**0.84**
PCIP[18]	0.87	0.86	0.73	0.78	0.83	0.62	0.54	0.65	0.74	0.85	0.83	**0.82**	0.51	**0.63**	0.81
TED[23]	0.89	**0.89**	—	**0.83**	—	—	—	—	—	—	—	—	—	—	—
CAPformer[24]	—	0.84	—	0.76	—	—	0.55	—	**0.76**	—	—	0.73	—	0.56	—
STFF-MANet[20]	0.89	0.88	0.79	0.82	0.85	—	—	—	—	—	—	—	—	—	—
Ours	**0.90**	0.87	**0.82**	0.81	0.80	**0.63**	**0.55**	**0.66**	0.75	**0.88**	**0.83**	0.80	**0.52**	0.61	0.79

以看出本文模型在PIE数据集上的ACC指标突破到了90%, Prec指标也有将近5%的提升, 在$JAAD_{beh}$和$JAAD_{all}$上也有多项指标优于现有模型.

3.3.2 消融实验

为了验证不同交叉注意力机制的实现方式对视觉特征融合的影响以及本文设计的融合注意力机制的分级GRU模块的有效性, 设计如表2所示的消融实验. 其中: CA(a)、CA(b)、CA(c)分别表示使用图2中3种交叉注意力进行视觉特征融合编码, CA-GRU表示本文提出的融合注意力机制的非视觉特征融合编码器, 除Ours方法使用了CA-GRU, 其他3种方法均使用文献[19]中的分级GRU进行编码. 由结果分析可知, 图2(c)中方式的交叉注意力实现方式能更好地对原始图像与语义图像之间进行信息交互.

表 2 不同模块消融实验对比

models	modules				PIE				
	CA(a)	CA(b)	CA(c)	CA-GRU	ACC	AUC	Prec	F1	Recall
Ours1	√				0.90	0.85	0.85	0.80	0.76
Ours2		√			0.89	0.87	0.83	0.77	0.80
Ours3			√		0.90	0.84	**0.85**	0.79	0.80
Ours			√	√	**0.90**	**0.87**	0.82	**0.81**	**0.80**

为了进一步验证CA-GRU的有效性, 控制视觉特征融合模块使用图2(c)不变, 对比已有的几种非视觉特征融合方式和本文所提出的CA-GRU融合方式的性能, 实验结果如表3所示, 使用CA-GRU进行非视觉特征的融合可以在多个指标上有一定提升.

表 3 CA-GRU与其他非视觉特征融合方式对比

models	PIE				
	ACC	AUC	Prec	F1	Recall
SingleGRU	0.83	0.77	0.70	0.67	0.64
SF-GRU	0.90	0.84	**0.85**	0.79	0.80
U-GRU	0.88	0.85	0.81	0.80	0.77
CA-GRU	**0.90**	**0.87**	0.82	**0.81**	**0.80**

另一方面, 本文还对不同的图像裁切方式进行实验, 对比了只使用检测框内图像(bbox)、使用整张图像(scene)、使用行人所在的四分之一图像(local), 结果如表4所示. 由表4结果可知, 行人所在位置的四分之一图像能更好地表达出行人外观及周围环境因素对行人过街行为的影响, 而只使用行人检测框内图像的话, 几乎没有任何环境信息, 使用整张图像又会掺杂过多无关环境信息(天空、背景建筑、树木等).

表 4 图像不同裁剪方式的消融实验结果

image	ACC	AUC	Prec	F1	Recall
bbox	0.87	0.83	0.72	0.78	0.77
scene	0.89	0.86	0.73	0.78	0.80
local	**0.90**	**0.87**	**0.82**	**0.81**	**0.80**

3.3.3 定性分析

图5展示了几个真实样本和其真实标签(GT)以及PCPA模型、PCIP模型、本文模型的预测结果. 图中4个场景均为交叉路口, 单独分析行人的运动特征很难预测行人接下来是否会发生过街行为, 但不同的是左侧两样本中交通信号灯为绿色或没有交通信号灯, 而右侧两样本场景的交通信号灯均为红色, 4个样本中均存在其他车辆. 图5中: C表示行人发生了过街行为(cross), NC表示没有发生过街行为(no cross).

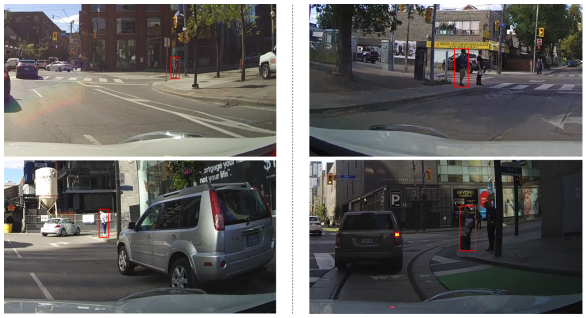

图 5　定性结果

由图5可知,本文方法在这些复杂的交叉路口场景下成功准确地预测了行人的过街意图,而另外两个基准模型均预测失败,表明本文方法对环境的特征提取与理解能力更强.

4　结　论

本文提出了一种基于混合注意力机制的多信息行人过街意图预测模型,通过交叉注意力实现原始图像特征与语义图像特征之间的信息交互,并设计了一种融合注意力机制的分级GRU模块提取行人骨骼关键点、历史轨迹等不同维度信息的运动特征,最终融合预测行人过街意图. 本文方法在PIE和JAAD数据集上的多项指标均优于现有其他模型. 未来可以对行人相对于环境是否真实存在运动进行研究,而不是简单使用驾驶视角下行人检测框的变化代表行人的运动,使模型更加接近于真人驾驶员的思维逻辑,进一步提高模型的性能与安全性.

参考文献(References)

[1] Rasouli A, Kotseruba I, Tsotsos J K. Pedestrian action anticipation using contextual feature fusion in stacked RNNs[C]. 30th British Machine Vision Conference. Cardiff: BMVA Press, 2020: 1-13.

[2] 刘建伟, 宋志妍. 循环神经网络研究综述[J]. 控制与决策, 2022, 37(11): 2753-2768.
(Liu J W, Song Z Y. Overview of recurrent neural networks[J]. Control and Decision, 2022, 37(11): 2753-2768.)

[3] Kotseruba I, Rasouli A, Tsotsos J K. Benchmark for evaluating pedestrian action prediction[C]. 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, 2021: 1257-1267.

[4] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks[C]. 2015 IEEE International Conference on Computer Vision. Santiago, 2015: 4489-4497.

[5] Saleh K, Hossny M, Nahavandi S. Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal DenseNet[C]. 2019 International Conference on Robotics and Automation. Montreal, 2019: 9704-9710.

[6] Rasouli A, Yau T, Rohani M, et al. Multi-modal hybrid architecture for pedestrian action prediction[C]. 2022 IEEE Intelligent Vehicles Symposium. Aachen, 2022: 91-97.

[7] Schneemann F, Heinemann P. Context-based detection of pedestrian crossing intention for autonomous driving in urban environments[C]. 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems. Daejeon, 2016: 2243-2248.

[8] Rasouli A, Kotseruba I, Tsotsos J K. Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior[C]. 2017 IEEE International Conference on Computer Vision Workshops. Venice, 2017: 206-213.

[9] Kotseruba I, Rasouli A, Tsotsos J K. Do they want to cross? Understanding pedestrian intention for behavior prediction[C]. 2020 IEEE Intelligent Vehicles Symposium. Las Vegas, 2020: 1688-1693.

[10] Lorenzo J, Parra I, Wirth F, et al. RNN-based pedestrian crossing prediction using activity and pose-related features[C]. 2020 IEEE Intelligent Vehicles Symposium. Las Vegas, 2020: 1801-1806.

[11] Fang Z J, López A M. Is the pedestrian going to cross? Answering by 2D Pose Estimation[C]. 2018 IEEE Intelligent Vehicles Symposium. Changshu, 2018: 1271-1276.

[12] Cadena P R G, Yang M, Qian Y Q, et al. Pedestrian graph: Pedestrian crossing prediction based on 2D pose estimation and graph convolutional networks[C]. 2019 IEEE Intelligent Transportation Systems Conference. Auckland, 2019: 2000-2005.

[13] Rasouli A, Kotseruba I, Kunic T, et al. PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction[C]. 2019 IEEE/CVF International Conference on Computer Vision. Seoul, 2019: 6261-6270.

[14] Piccoli F, Balakrishnan R, Perez M J, et al. FuSSI-Net: Fusion of spatio-temporal skeletons for intention prediction network[C]. 2020 54th Asilomar Conference on Signals, Systems, and Computers. Pacific Grove, 2020: 68-72.

[15] Saleh K, Hossny M, Nahavandi S. Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal DenseNet[C]. 2019 International Conference on Robotics and Automation. Montreal, 2019: 9704-9710.

[16] Saleh K, Hossny M, Nahavandi S. Spatio-temporal DenseNet for real-time intent prediction of pedestrians in urban traffic environments[J]. Neurocomputing, 2020, 386: 317-324.

[17] 杨彪, 范福成, 杨吉成, 等. 基于动作预测与环境条件的行人过街意图识别[J]. 汽车工程, 2021, 43(7): 1066-1076.
(Yang B, Fan F C, Yang J C, et al. Recognition of pedestrians' street-crossing intentions based on action prediction and environment context[J]. Automotive Engineering, 2021, 43(7): 1066-1076.)

[18] Yang D F, Zhang H L, Yurtsever E, et al. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention[J]. IEEE Transactions on Intelligent Vehicles, 2022, 7(2): 221-230.

[19] Rasouli A, Rohani M, Luo J. Pedestrian behavior prediction via multitask learning and categorical interaction modeling[C]. 2022 International Conference on Robotics and Automation. Pjiladelphia, 2022: 940-947.

[20] Zhang X F, Wang X L, Zhang W W, et al. Multi-attention network for pedestrian intention prediction based on spatio-temporal feature fusion[J]. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, https://doi.org/10.1177/09544070231190522.

[21] Guo D X, Mordan T, Alahi A. Pedestrian stop and go forecasting with hybrid feature fusion[C]. 2022 International Conference on Robotics and Automation (ICRA). New York: ACM, 2022: 940-947.

[22] Gesnouin J, Pechberti S, Stanciulcscu B, et al. TrouSPI-Net: Spatio-temporal attention on parallel atrous convolutions and U-GRUs for skeletal pedestrian crossing prediction[C]. 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). New York: ACM, 2021: 1-7.

[23] Upreti M, Ramesh J, Kumar C, et al. Traffic light and uncertainty aware pedestrian crossing intention prediction for automated vehicles[C]. 2023 IEEE Intelligent Vehicles Symposium. Anchorage, 2023: 1-8.

[24] Vaswani A, Shazeer N, Parmar N, et al. Attentionis all you need[C]. Advances in Neural Information Processing Systems. Long Beach: NIPS Foundation, 2017: 5998-6008.

[25] Lorenzo J, Alonso I P, Izquierdo R, et al. CAPformer: Pedestrian crossing action prediction using transformer[J]. Sensors, 2021, 21(17): 5694.

[26] Fan L X, Buch S, Wang G Z, et al. RubiksNet: Learnable 3D-shift for efficient video action recognition[C]. European Conference on Computer Vision. Cham: Springer, 2020: 505-521.

[27] 侯志强, 郭凡, 杨晓麟, 等. 基于混合注意力的Transformer视觉目标跟踪算法[J]. 控制与决策, 2024, 39(3): 739-748.
(Hou Z Q, Guo F, Yang X L, et al. Transformer visual object tracking algorithm based on mixed attention[J]. Control and Decision, 2024, 39(3): 739-748.)

作者简介

桑海峰(1978−), 男, 教授, 博士, 博士生导师, 主要研究方向为机器视觉检测和智能视频分析, E-mail: sanghaif@163.com;

刘玉龙(2000−), 男, 硕士生, 主要研究方向为行人过街意图预测, E-mail: liuyulong000301@163.com;

刘泉恺(1998−), 男, 硕士生, 主要研究方向为行人轨迹预测, E-mail: liuqk_sut@qq.com.