基于第一视角的非自回归行人轨迹预测模型

桑海峰,王金玉,陈旺兴,王海峰
(沈阳工业大学信息科学与工程学院,辽宁沈阳 110870)

摘 要: 行人轨迹预测在自动驾驶和监控系统等多个应用中具有重要意义.目前大多数行人轨迹预测模型采用基于循环神经网络的编码器-解码器结构,其自回归的解码结构存在一定的累积误差,而且循环神经网络对序列的长期依赖问题仍然无法很好地解决.本文提出一种基于Transformer网络的非自回归行人轨迹预测模型,非自回归的解码结构能够同时生成所有预测值来减少累积误差,Transformer网络中的自注意力机制能够改善长期依赖问题.本文还设计一个局部信息加强模块来捕获行人运动趋势发生变化的局部特征,同时结合边界框的位置信息和大小信息来编码第一视角下透视投影产生的影响,使得模型提取到的轨迹特征更加有效.实验结果表明,在基于第一视角的公开数据集PIE(Pedestrian Intention Estimation)上,本文提出的模型比PIE预测模型在15、30、45帧的平均位移误差和终点位移误差上分别降低了24%,14.5%,11%和6%.

关键词: 行人轨迹预测;第一视角;Transformer网络;非自回归预测;累积误差;局部信息加强
基金项目: 国家自然科学基金(No.62173078);辽宁省教育厅科研项目(No.LJGD2020006)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2023)05-1266-07
电子学报URL: http://www. ejournal. org. cn **DOI:** 10.12263/DZXB.20211467

Non-Autoregressive Pedestrian Trajectory Prediction Model Based on the First Perspective

SANG Hai-feng, WANG Jin-yu, CHEN Wang-xing, WANG Hai-feng
(*School of Information Science and Engineering, Shenyang University of Technology, Shenyang, Liaoning* 110870, *China*)

Abstract: Pedestrian trajectory prediction plays an important role in many applications such as automatic driving and monitoring systems. At present, most pedestrian trajectory prediction models are recurrent neural network (RNN) based on encoder-decoder architectures. RNN could not solve the long-term dependence, and its auto-regressive decoding scheme introduces accumulate errors. This paper proposes a Transformer based non-autoregressive pedestrian trajectory prediction model, whose non-autoregressive decoder can generate all predictions simultaneously to reduce accumulative errors. The self-attention mechanism can enhance the long-term dependence problem. More specifically, this paper designs a local information enhancement module to extract the local features when pedestrian's movement trend changes, and combining with the location information and scale of the boundary encodes the impact of perspective projection in the first perspective, which makes the trajectory features extracted from the model more efficient. Experimental results show that, compared with the PIE (Pedestrian Intention Estimation) model, the average displacement error of 15, 30 and 45 frame and the end displacement error are respectively reduced by 24%,14.5%,11% and 6% on a public data set PIE based on the first perspective.

Key words: pedestrian trajectory prediction; the first perspective; Transformer; non-autoregressive prediction; accumulative errors; local information enhancement module

Foundation Item(s): National Natural Science Foundation of China (No.62173078); Research Project of Liaoning Education Department (No.LJGD2020006)

1 引言

行人轨迹预测在自动驾驶和监控系统等多个应用中具有重要意义,也是计算机视觉领域中具有挑战性的任务之一.轨迹预测任务基于已有的真实轨迹信息来预测未来一段时间的轨迹,通常被描述为序列建模问题.因此,建立有效的序列模型来捕捉行人轨迹之间

收稿日期:2021-10-29;修回日期:2021-12-27;责任编辑:覃怀银

的时间依赖性是解决这一问题的关键. 随着深度学习的快速发展,基于循环神经网络的模型,如长短期记忆网络(Long Short-Term Memory,LSTM)[1]或门控循环单元(Gated Recurrent Unit,GRU)[2],在序列建模中取得了巨大的成功.

目前大多数轨迹预测模型通常是解码器-编码器结构以自回归的方式预测未来轨迹[3~10],这种自回归的预测模型通常会遇到一个问题——累积误差,尽管LSTM网络在一定程度上缓解了长期依赖性问题,但是对于特别长的序列,仍然不能很好的解决该问题. 最近一段时间,Transformer网络[11]受到了各界的关注,其强大的自注意力机制在计算不同位置之间的相似性时采用点对点的计算方式,缓解长距离依赖问题. Giuliari F等人[12]首先将Transformer网络应用到监控场景下的行人轨迹预测中,对每个行人单独建模,在没有考虑任何人-人或者人-场景交互的情况下取得了非常好的效果. 但是Transformer网络的解码器也是自回归结构,仍然存在一定的累积误差. Li等人[13]在动作序列预测中提出了非自回归Transformer模型,受文献[13]启发,本文设计一个非自回归构造模块用来生成隐向量,该向量作为解码器的输入对每个时间点的轨迹数据独立预测,从而缓解由自回归解码机制产生的累积误差问题.

由于行人的运动具有随机性与不确定性,行人会根据所处环境的改变而随时改变自己的运动趋势,例如,在遇到静态障碍物时,可能会绕开再继续行走;在遇到车辆时,可能会减速行走以避开车辆,也可能会加速前进;在十字路口时,可能会先停下来观望然后再继续行走. Transformer网络虽能很好的捕获全局特征,学习行人的整体运动趋势,然而在运动趋势发生改变时,自注意力机制可能会忽略这种局部信息,按照之前的运动趋势进行预测必然会生成错误的轨迹. 判断一个时间点的运动趋势是否发生变化,往往需要根据其上下文来确定[14],本文设计了一个局部信息加强模块来关注运动趋势发生改变时的信息,提升轨迹预测的准确性.

在第一视角下,行人在图像中移动的距离与现实世界中的物理距离并不直接对应,同时由于透视投影的原因,行人在图像中会出现“近大远小”的情况,只提取行人的位置信息难以准确的预测未来轨迹[15~17],本文同时考虑了行人的位置信息和边界框的大小信息来编码轨迹特征.

2　基于Transformer的非自回归行人轨迹预测模型

本文提出一种基于Transformer的非自回归行人轨迹预测模型,结构框图如图1所示. 轨迹序列输入到编码器之前,先经过一个局部信息加强模块提取轨迹序列中运动趋势发生变化的局部特征,然后再经过编码器编码,既能提取全局特征也能有效捕获局部特征. 非自回归构造模块一次性生成所有能表示目标序列相关性的隐变量,解码器对隐变量解码,同时预测出所有时间点的轨迹值,从而减少自回归预测模型中存在的累积误差.

2.1　问题描述

行人轨迹预测任务可以表示为对场景中任意行人,观察他在过去一段时间内的历史轨迹来预测其未来轨迹. 对于行人i,假设他在时间t的位置向量为$\boldsymbol{l}_t^i$,行人i的观测轨迹为$(\boldsymbol{l}_{t-t_{\text{obs}}+1}^i,\cdots,\boldsymbol{l}_{t-1}^i,\boldsymbol{l}_t^i)$,预测轨迹为$(\boldsymbol{l}_{t+1}^i,\boldsymbol{l}_{t+2}^i,\cdots,\boldsymbol{l}_{t+t_{\text{pred}}}^i)$. 本文的目标是从当前第$t$帧开始预测未来$t_{\text{pred}}$帧相对第一帧的位置,即观测轨迹$(\boldsymbol{l}_{t-t_{\text{obs}}+2}^i-\boldsymbol{l}_{t-t_{\text{obs}}+1}^i,\cdots,\boldsymbol{l}_{t-1}^i-\boldsymbol{l}_{t-t_{\text{obs}}+1}^i,\boldsymbol{l}_t^i-\boldsymbol{l}_{t-t_{\text{obs}}+1}^i)$,预测轨迹为$(\boldsymbol{l}_{t+1}^i-\boldsymbol{l}_{t-t_{\text{obs}}+1}^i,\boldsymbol{l}_{t+2}^i-\boldsymbol{l}_{t-t_{\text{obs}}+1}^i,\cdots,\boldsymbol{l}_{t+t_{\text{pred}}}^i-\boldsymbol{l}_{t-t_{\text{obs}}+1}^i)$,其中,$t_{\text{obs}}$为观测时间,$t_{\text{pred}}$为预测时间.

2.2　模型输入

在第一视角下,视频中行人的移动距离与现实世界中行人的移动距离不直接对应,例如,位于图像中心的行人可能在车辆附近,也可能在远处过街,这种差异会导致截然不同的未来轨迹. 因此本文结合行人边界框的位置坐标和大小信息一起编码轨迹序列,对于行人i在时间t的输入向量表示为$\boldsymbol{l}_t^i=\left(x_t^i,y_t^i,w_t^i,h_t^i\right)$,其中,$\left(x_t^i,y_t^i\right)$表示行人$i$在$t$时刻边界框的中心坐标点,$\left(w_t^i,h_t^i\right)$表示$t$时刻边界框的大小信息.

2.3　位置编码器

循环神经网络本身是一种顺序结构,天生包含序列的位置信息. 当Transformer网络抛弃循环网络结构,采用注意力机制取而代之,丢失了原本序列中的位置信息. 位置编码模块将生成的位置向量加到输入向量上帮助模型学习位置信息,给定一个长度为n的序列,x表示数据在序列中的位置,位置编码器定义如式(1)所示.

$$\begin{cases}\text{pos}(x,2i)=\sin\left(\dfrac{x}{10\,000^{2i/d_{\text{model}}}}\right)\\ \text{pos}(x,2i+1)=\cos\left(\dfrac{x}{10\,000^{2i+1/d_{\text{model}}}}\right)\end{cases}\tag{1}$$

其中,$\text{pos}(x,2i)$,$\text{pos}(x,2i+1)$表示x处位置的偶数维度编码和奇数维度编码,d_{model}表示向量维度.

由正弦函数生成的位置编码具有连续性与高度相关性,即位置x_1,x_2越接近,$\text{pos}(x_1)$和$\text{pos}(x_2)$越相似,对于任意固定的偏移量δ,$\text{pos}(x+\delta)$可以表示成$\text{pos}(x)$的线性函数.

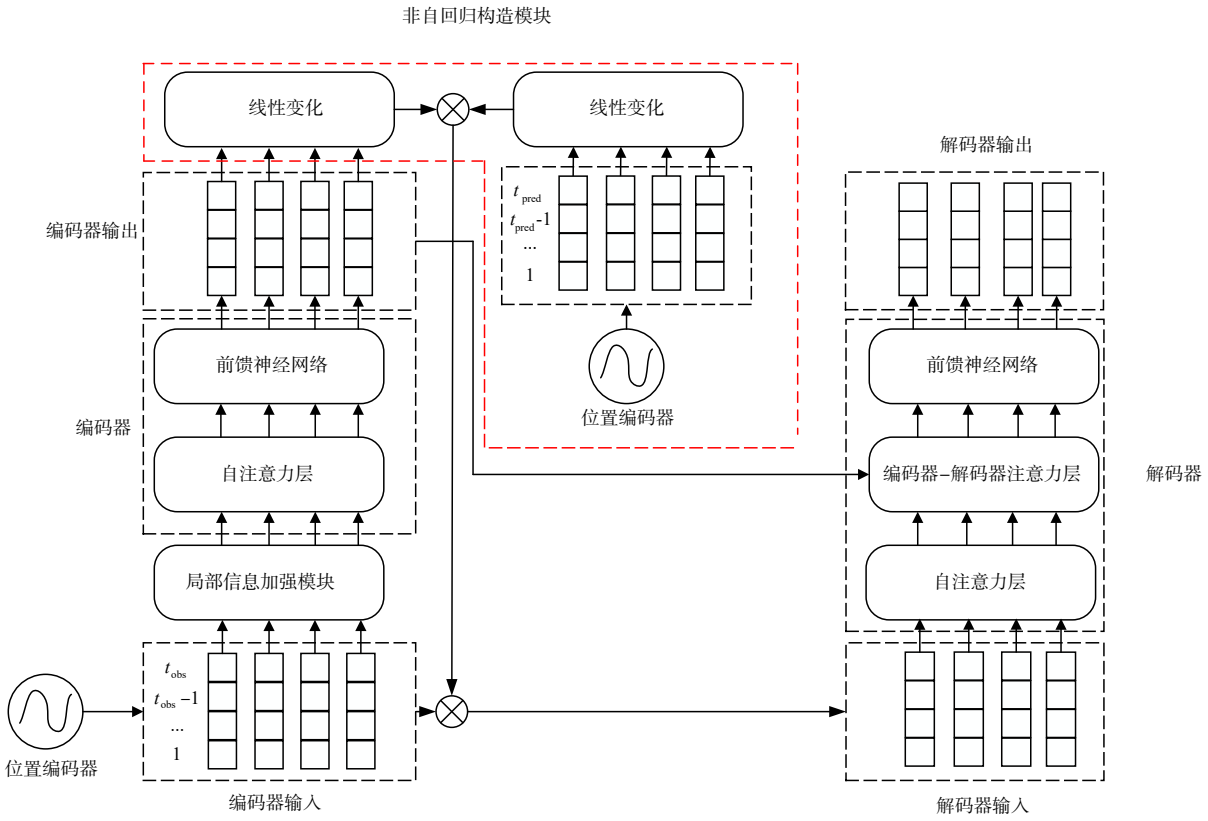


图1 基于Transformer的非自回归行人轨迹预测模型结构框图

2.4 局部信息加强模块

在行人轨迹预测中,若行人以恒定的速度运动,自注意力机制能够很好地学习到行人的运动趋势. 在现实场景中,行人会随着场景的变化做出不同的决策,随时改变运动方向,如图2所示,图中每个点表示行人边界框中心点在图片中的位置. 若解码器解码时再继续关注运动趋势改变之前的轨迹信息必然会产生错误的预测,因此本文提出一个局部信息加强模块来提取观测轨迹中的局部特征.

局部信息加强模块首先将行人轨迹序列整合成矩阵形式,如图3所示. 我们可以把矩阵看成是一幅图像,使用一维卷积网络按着垂直方向滑动提取特征,步长为1,卷积核的宽度为轨迹特征的维度. 经过卷积处理之后的输入向量更加关注序列中运动趋势发生变化的局部特征,然后再输入到编码器中提取全局特征. 局部信息加强模块示意图如图4所示.

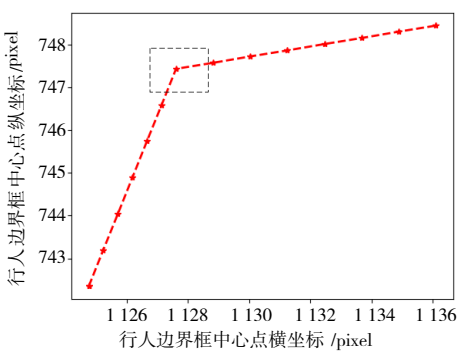


图2 运动趋势变化图

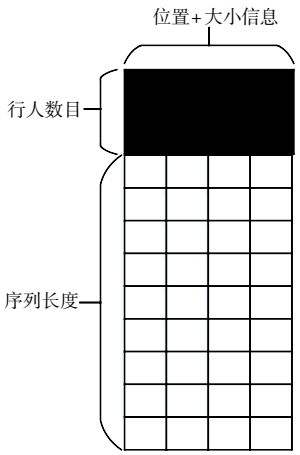


图3 轨迹序列矩阵示意图

局部信息加强模块对行人i的观测轨迹$\boldsymbol{X}^i$采用一维卷积进行局部特征提取,对运动趋势发生变化的时间点赋予较大的权重,计算公式如式(2)所示.

$$\hat{\boldsymbol{X}}^i=f_{\mathrm{CNN}}\left(\boldsymbol{X}^i,\boldsymbol{W}_{\mathrm{CNN}}\right) \tag{2}$$

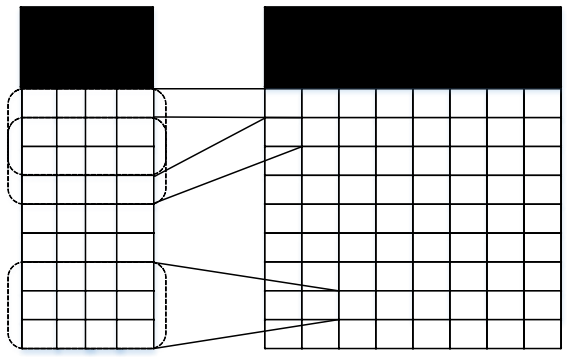


图4　局部信息加强模块示意图

其中,$\boldsymbol{W}_{\mathrm{CNN}}$为可学习的参数矩阵,$\boldsymbol{X}^i$为模块输入向量,$\hat{\boldsymbol{X}}^i$为模块的输出向量,$f_{\mathrm{CNN}}$为一维卷积函数.

2.5　非自回归 Transformer

2.5.1　编码器

编码器部分与原始Transformer编码器一致,每个编码层都是由自注意力层和前馈神经网络层组成. 编码器的输入向量首先会经过自注意力层如式(3)所示,该层帮助编码器计算当前位置和序列中其他位置之间的相似性,自注意力层的输出再传递到前馈神经网络.

$$\begin{aligned}&\boldsymbol{q}=\hat{\boldsymbol{X}}\boldsymbol{W}_q\\&\boldsymbol{k}=\hat{\boldsymbol{X}}\boldsymbol{W}_k\\&\boldsymbol{v}=\hat{\boldsymbol{X}}\boldsymbol{W}_v\\&\mathrm{attention}\left(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}\right)=\mathrm{softmax}\left(\frac{\boldsymbol{q}\boldsymbol{k}^{\mathrm{T}}}{\sqrt{d_k}}\right)\end{aligned}\tag{3}$$

其中,$\boldsymbol{W}_q,\boldsymbol{W}_k,\boldsymbol{W}_v$为权重矩阵,$\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}$分别为查询向量,键向量和值向量,$\hat{\boldsymbol{X}}$为自注意力层输入矩阵,$d_k$为权重矩阵维度.

2.5.2　非自回归解码器

在行人轨迹预测中,t时刻的位置是根据前$t-1$时刻的位置进行预测,如式(4)所示,在预测前$t-1$时刻时,不可避免会产生一定的误差,在进行后面的预测时误差逐渐累积,导致最终的结果不理想,同时这种模型也不能并行训练. 本文构造一个非自回归解码器,使其在解码时能够同时生成整个序列,既能够减少累积误差,又能使模型并行训练.

由于目标序列具有时空相关性,因此需要引入一个隐向量$\boldsymbol{Z}$(即解码器输入向量),如式(5)所示. 在给定$\boldsymbol{Z}$的情况下,由于目标序列中每个数据都是相互独立的,则存在$p\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z};\theta\right)=\prod_{t=1}^{t_{\mathrm{pred}}}p\left(y_t|\boldsymbol{X},\boldsymbol{Z};\theta\right)$,得到式(6).

$$p_{\mathrm{AR}}\left(\boldsymbol{Y}|\boldsymbol{X};\theta\right)=\prod_{t=1}^{t_{\mathrm{pred}}}p\left(y_t|y_{0:t-1},\boldsymbol{X};\theta\right)\tag{4}$$

$$p_{\mathrm{NAR}}\left(\boldsymbol{Y}|\boldsymbol{X};\theta\right)=\int_z p\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z};\theta\right)p_{\mathrm{L}}\left(\boldsymbol{Z}|\boldsymbol{X};\theta\right)\mathrm{d}z\tag{5}$$

$$p_{\mathrm{NAR}}\left(\boldsymbol{Y}|\boldsymbol{X};\theta\right)=p_{\mathrm{L}}\left(\boldsymbol{Z}|\boldsymbol{X};\theta\right)\prod_{t=1}^{t_{\mathrm{pred}}}p\left(y_t|\boldsymbol{X},\boldsymbol{Z};\theta\right)\tag{6}$$

其中,AR表示自回归模型,NAR表示非自回归模型,p为条件概率分布函数,$\boldsymbol{X}$为观测序列,$\boldsymbol{Y}$为目标序列,$\boldsymbol{Z}$为隐向量,t_{pred}为预测时间,θ为可训练参数.

除了输入向量有差异,解码器的其余部分与原始Transformer的解码器一致,包含自注意力层,编码器-解码器注意力层和前馈神经网络. 自注意力层和前馈神经网络层与编码器中的一致,编码器-解码器注意力层类似于基于注意力机制的序列到序列模型,在预测当前时间步时,模型更加关注所有与当前预测相关的观测序列数据,使模型能够更准确的预测,减少由长序列造成的信息丢失问题.

2.5.3　非自回归构造模块

本文设计一个非自回归构造模块生成隐变量,将隐变量作为解码器的输入来生成每一时刻的预测值. 该模块以编码器的输入向量、输出向量和隐变量的位置编码向量作为输入,如图1红色框内所示. 由于本文预测的轨迹序列与观测序列长度不同,直接对编码器输入向量进行变换,导致隐变量的长度与预测长度不一致,因此采用线性变换g将隐变量的长度与预测长度对齐,为使隐变量包含更多特征信息,采用线性变换f将编码器的编码特征整合进隐变量中.

具体来说,在进行轨迹预测时,预测轨迹长度是事先知道的,得到隐向量的位置信息$\boldsymbol{D}_{\mathrm{pos}}^i\in\mathbb{R}^{t_{\mathrm{pred}}\times d_{\mathrm{model}}}$,将位置编码向量进行线性变换得到输出$\boldsymbol{P}^i\in\mathbb{R}^{t_{\mathrm{pred}}\times 1}$. 编码器的输出向量为$\boldsymbol{E}_{\mathrm{out}}^i\in\mathbb{R}^{t_{\mathrm{obs}}\times d_{\mathrm{model}}}$也进行一次线性变换得到输出为$\boldsymbol{Q}^i\in\mathbb{R}^{t_{\mathrm{obs}}\times 1}$. 结合以上两个信息,便得到隐变量$\boldsymbol{Z}^i\in\mathbb{R}^{t_{\mathrm{pred}}\times d_{\mathrm{model}}}$,然后输入到解码器中解码,见式(7)~(9)所示.

$$\boldsymbol{Q}^i=f\left(\boldsymbol{E}_{\mathrm{out}}^i\boldsymbol{W}_{\boldsymbol{Q}}+b_{\boldsymbol{Q}}\right)\tag{7}$$

$$\boldsymbol{P}^i=g\left(\boldsymbol{D}_{\mathrm{pos}}^i\boldsymbol{W}_{\boldsymbol{P}}+b_{\boldsymbol{P}}\right)\tag{8}$$

$$\boldsymbol{Z}^i=\boldsymbol{P}^i\left(\boldsymbol{Q}^i\right)^{\mathrm{T}}\boldsymbol{X}^i\tag{9}$$

其中,$\boldsymbol{X}^i\in\mathbb{R}^{t_{\mathrm{obs}}\times d_{\mathrm{model}}}$为行人$i$的编码器输入向量,$t_{\mathrm{obs}}$为观测时间,$t_{\mathrm{pred}}$为预测时间,$d_{\mathrm{model}}$为嵌入的维度,$\boldsymbol{W}_{\boldsymbol{P}},\boldsymbol{W}_{\boldsymbol{Q}}$,为可训练参数矩阵,$b_{\boldsymbol{P}},b_{\boldsymbol{Q}}$为可训练参数.

3　实验结果与分析

3.1　数据集和评价指标

该模型在一个基于第一视角的行人轨迹预测数据集(Pedestrian Intention Estimation,PIE)[10]上进行训练. PIE数据集由6个小时的驾驶素材组成,不仅包括人流

量高且道路狭窄的城市场景还包括行人较少且道路宽敞的街道. 在本文中,模型学习前15帧的轨迹信息,来预测后15帧、30帧、45帧的轨迹信息.

本文采用两种评价指标:

平均位移误差(Average Displacement Error, ADE):在预测的所有步长内,真实值与预测值的均方误差,如式(10)所示.

$$\mathrm{ADE}=\frac{\sum i\sum_{t=1}^{t_{\mathrm{pred}}}\left\|\hat{\boldsymbol{l}}_t^i-\boldsymbol{l}_t^i\right\|^2}{(\sum i)\times t_{\mathrm{pred}}} \tag{10}$$

其中,t_{pred}为预测时间,i为行人数目,$\boldsymbol{l}_t^i$为行人i在t时刻的真实位置向量,$\hat{\boldsymbol{l}}_t^i$为行人i在t时刻的预测位置向量.

终点位移误差(Final Displacement Error, FDE):预测的最后一个时间点的真实值与预测值的均方误差,如式(11)所示.

$$\mathrm{FDE}=\frac{\sum i\left\|\hat{\boldsymbol{l}}_t^i-\boldsymbol{l}_t^i\right\|^2}{\sum i},t=t_{\mathrm{pred}} \tag{11}$$

其中,t_{pred}为预测时间,i为行人数目,$\boldsymbol{l}_t^i$为行人i在t时刻的真实位置向量,$\hat{\boldsymbol{l}}_t^i$为行人i在t时刻的预测位置向量.

本文模型将与以下模型进行对比,以下为各模型的简单介绍:

Linear[18]:线性卡尔曼滤波模型

LSTM[1]:原始LSTM网络

B_LSTM[19]:基于贝叶斯估计的LSTM模型

$\mathrm{PIE}_{\mathrm{traj}}$[10]:基于LSTM的序列-序列模型,引入时间注意力机制和自注意力机制,前者捕获关键帧信息,后者捕获与当前预测相关的特征级信息

3.2 模型构造和训练设置

模型设置采用了原始Transformer网络的设置,d_{model}=512,编码器和解码器堆叠了6层. 局部信息加强模块中,卷积核大小为4,步长为1. 在训练过程中,训练批次大小设置为128,采用随机优化方法,设置自动衰减学习率策略,初始学习率为0.000 1. 本文的实验均在相同的硬件条件下进行,处理器Intel(R) Core(TM) i7-8700 CPU @ 3.20 GHz内核数量为6,显卡NVIDIA 2080TI GPU显存为11 GB.

3.3 定量分析

3.3.1 模型预测结果对比

在15帧、30帧和45帧的平均位移误差和45帧的终点位移误差指标上比较了本文模型与其他模型的预测结果对比,具体结果见表1,其中ADE_15、ADE_30、ADE_45分别表示预测15帧、30帧、45帧的平均位移误差,FDE为45帧的终点位移误差. 本文提出的模型在所有指标上均优于上面所提到的模型,本文模型比PIE的轨迹预测模型在各个指标上分别下降了24%,14.5%,11%和6%.

表1 本文模型与其他经典模型的预测结果对比表 单位:pixel

	ADE_15	ADE_30	ADE_45	FDE
Linear	123	477	1 365	3 983
LSTM	172	330	911	3 352
B-LSTM	101	296	855	3 259
$\mathrm{PIE}_{\mathrm{traj}}$	58	200	636	2 569
本文模型	44	171	568	2 404

3.3.2 自回归模型和非自回归模型预测结果对比

非自回归的Transformer网络在ADE_15、ADE_30、ADE_45和FDE指标上分别降低了24%,45%,53%和57%,尤其是在预测长度变长时,预测结果提升明显. 表2为基于Transformer网络的自回归模型与非自回归模型的预测结果对比表.

表2 自回归与非自回归模型预测结果对比表 单位:pixel

	ADE_15	ADE_30	ADE_45	FDE
自回归	94	447	1 610	7 191
非自回归	71	242	752	3 058

3.3.3 有无局部信息加强模块的实验对比

表3为有无局部信息加强模块的预测结果对比表,无为未添加局部信息加强模块,有为添加局部信息加强模块. 实验结果表明添加了局部信息加强模块的预测效果较好,在ADE_15、ADE_30、ADE_45和FDE上分别降低了17%,9.5%,5.7%和4.7%.

表3 有无局部信息加强模块的预测结果对比表 单位:pixel

	ADE_15	ADE_30	ADE_45	FDE
无	71	242	752	3 058
有	59	219	709	2 918

3.3.4 模型不同输入的实验对比

表4为有无边界框大小信息的预测结果对比表,无表示没有加入边界框大小信息,有表示加入边界框大小信息. 如表4所示,增加了边界框大小信息的预测结果要优于只利用位置信息的预测结果.

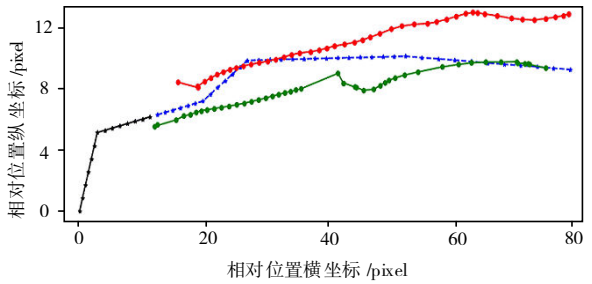
表4 模型不同输入的预测结果对比表 单位:pixel

	ADE_15	ADE_30	ADE_45	FDE
无	86	261	782	3 113
有	71	242	752	3 058

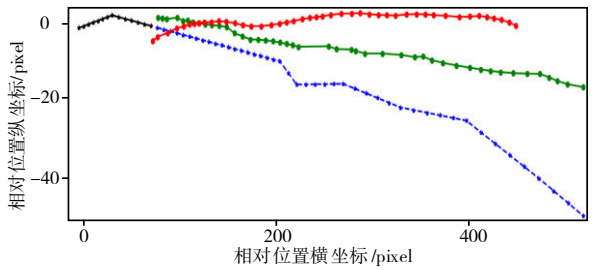
3.3.5 预测结果对比

图5为本文的模型与PIE模型的预测轨迹对比图,根据15帧的观测轨迹来预测未来45帧的轨迹,其中黑

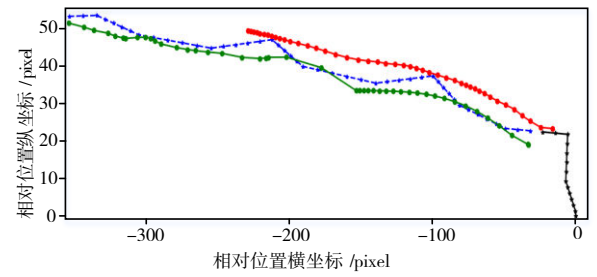
色代表的是观测轨迹,蓝色为真实轨迹,红色为PIE模型的预测轨迹,绿色为本文模型的预测轨迹. 图5中每一点表示当前帧行人边界框中心点与第一帧行人边界框中心点的相对位置,可以明显看出本文的预测模型更加接近真实轨迹.



(a) 行人1



(b) 行人2



(c) 行人3

图5　本文模型与PIE_{traj}模型的预测轨迹对比图

3.4　定性分析

在实验对比部分可以看出本文提出的模型在评价指标上要优于之前的模型. 现在定性的分析,本文提出的非自回归模型为什么能减少累积误差,考虑了轨迹序列局部特征为何能提升预测效果.

以自回归的方式进行轨迹预测时,下一帧的预测是根据观测到的真实轨迹和当前的预测值生成的,之前帧的预测难免出现误差,随着预测长度的增加,这种误差会累积的越来越大,导致最终的预测结果不理想. 本文提出的非自回归模型打破了这种串行顺序,尝试同时预测一整个序列,从而解决上述问题. 本文模型中的非自回归构造模块一次性生成所有能表达目标序列相关性的输入向量,从而构造一个非自回归的解码器同时预测出所有值. 行人在运动时会根据他所处的环境做出决策随时改变自己的轨迹,若根据运动趋势发生变化之前的信息进行预测必然会产生错误的轨迹,而一个时间点是否是运动趋势发生变化的点取决于他的上下文环境. 自注意力机制虽能很好地提取全局特征,但却忽略了局部特征,本文提出的局部信息加强模块对输入的轨迹序列先进行局部特征提取,然后再输入到编码器中提取全局特征. 如图5所示,相比较于PIE模型,本文提出的模型更接近于真实轨迹,可以看到PIE模型在预测长度变长时由于存在累积误差,使得预测轨迹越来越偏离真实轨迹,难以捕捉序列的长期依赖性,本文提出的更加关注局部信息的基于Transformer非自回归预测模型能够较好地学习到真实轨迹的特征,不仅改善了长期依赖问题也在一定程度上消除了累积误差的问题.

4　结论

本文提出了一种基于Transformer的非自回归轨迹预测模型,在PIE数据集上的轨迹预测结果较好. 本文模型结合了轨迹位置信息和边界框的大小信息,更好的编码行人轨迹序列,然后通过一个局部信息加强模块提取序列中的局部特征,关注运动趋势发生变化的时间点,再传递到自注意力层中提取序列的全局特征. 非自回归构造模块生成一个能表达目标序列相关性的输入向量,构造非自回归解码器同时预测出所有轨迹,本文提出的模型对于将来基于第一视角的行人轨迹预测提供了新思路.

参考文献

[1] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[2] CHO K, MERRIENBOER B VAN, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014). Stroudsburg: Association for Computational Linguistics, 2014: 1724-1734.

[3] ALAHI A, GOEL K, RAMANATHAN V, et al. Social LSTM: human trajectory prediction in crowded spaces[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 961-971.

[4] ZHANG P, OUYANG W L, ZHANG P F, et al. SR-LSTM: State refinement for LSTM towards pedestrian trajectory prediction[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 12077-12086.

[5] XUE H, HUYNH D Q, REYNOLDS M. SS-LSTM: A hi-

erarchical LSTM model for pedestrian trajectory prediction [C]//2018 IEEE Winter Conference on Applications of Computer Vision(WACV). Piscataway: IEEE, 2018: 1186-1194.
[6] LI J C, MA H B, TOMIZUKA M. Conditional generative neural system for probabilistic trajectory prediction[C]// 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE, 2020: 6150-6156.
[7] XUE H, HUYNH D Q, REYNOLDS M. A location-velocity-temporal attention LSTM model for pedestrian trajectory prediction[J]. IEEE Access, 2020, (8): 44576-44589.
[8] LIANG J W, JIANG L, NIEBLES J C, et al. Peeking into the future: Predicting future person activities and locations in videos[C]//2019 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 5718-5727.
[9] CUI H G, RADOSAVLJEVIC V, CHOU F C, et al. Multimodal trajectory predictions for autonomous driving using deep convolutional networks[C]//2019 International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2019: 2090-2096.
[10] RASOULI A, KOTSERUBA I, KUNIC T, et al. PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction[C]//2019 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2020: 6261-6270.
[11] GIULIARI F, HASAN I, CRISTANI M, et al. Transformer networks for trajectory forecasting[C]//2020 25th International Conference on Pattern Recognition (ICPR). Piscataway: IEEE, 2021: 10335-10342.
[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2017: 5998-6008.
[13] LI B, TIAN J, ZHANG Z F, et al. Multitask non-autoregressive model for human motion prediction[J]. IEEE Transactions on Image Processing, 2020, 30: 2562-2574.
[14] LI S, JIN X, XUAN Y, et al. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting[J]. Advances in Neural Information Processing Systems, 2019, (32): 5243-5253.
[15] YAO Y, XU M Z, CHOI C, et al. Egocentric vision-based future vehicle localization for intelligent driving assistance systems[C]//2019 International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2019: 9711-9717.
[16] YAGI T, MANGALAM K, YONETANI R, et al. Future person localization in first-person videos[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7593-7602.
[17] 白萧. 第一视角下的行人轨迹预测方法研究[D]. 大连: 大连海事大学, 2020.
[18] KALMAN R E. A new approach to linear filtering and prediction problems[J]. Journal of Basic Engineering, 1960, 82(1): 35-45.
[19] BHATTACHARYYA A, FRITZ M, SCHIELE B. Long-term on-board prediction of people in traffic scenes under uncertainty[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 4194-4202.

作者简介

桑海峰 男,1978年1月出生于辽宁省沈阳市.现为沈阳工业大学视觉检测研究所教授、博士生导师.主要研究方向为机器视觉检测技术和智能视频分析技术.
E-mail: sanghaif@163.com

王金玉(通讯作者) 女,1996年5月出生于辽宁省盖州市.现为沈阳工业大学视觉检测研究所博士.主要研究方向为行人轨迹预测.
E-mail: 1911131982@qq.com

陈旺兴 男,1998年3月出生于江西省抚州市.现为沈阳工业大学视觉检测研究所硕士.主要研究方向为行人轨迹预测.
E-mail: 1909703861@qq.com

王海峰 男,1995年3月出生于吉林省吉林市.现为沈阳工业大学视觉检测研究所硕士.主要研究方向为行人轨迹预测.
E-mail: 798466420@qq.com