doi:10.19562/j.chinasae.qcgc.2024.07.009

引入自监督预训练的轨迹预测方法*

李琳辉,付一帆,王 霆,王雪成,连 静
(大连理工大学机械工程学院,大连 116024)

[摘要] 针对目前基于监督学习的轨迹预测模型数据利用效率低、精度有限的问题,提出一种轨迹预测模型及通用的自监督预训练策略。首先,基于Transformer搭建轻量化的轨迹预测模型,实现场景时序空间特征提取与交互关系建模;其次,设计运动信息时序掩码、道路信息空间掩码、交互关系掩码3类掩码重建任务对模型进行自监督预训练,以提升模型对场景通用特征的提取能力;最后,以预训练权重为初始化参数在下游任务中进行监督学习微调。在Argoverse2 Motion Forecasting数据集的实验表明,模型在预训练任务中能够很好地重建出交通场景,引入自监督预训练能够有效提升预测精度和数据利用效率,且对不同预测任务具有通用性,在单目标轨迹预测与多目标轨迹预测任务上$minFDE_6$指标分别提升3.3%与3.7%。

关键词:自动驾驶;轨迹预测;自监督预训练

Trajectory Prediction Method Enhanced by Self-supervised Pretraining

Li Linhui, Fu Yifan, Wang Ting, Wang Xuecheng & Lian Jing
School of Mechanical Engineering, Dalian University of Technology, Dalian 116024

[**Abstract**] To address limitation in prediction accuracy and data utilization efficiency of supervised learning-based trajectory prediction models, a trajectory prediction model and a general self-supervised pretraining strategy are proposed. Firstly, a lightweight trajectory prediction model based on Transformer is established to extract temporal-spatial features while modeling interaction relationship. Secondly, three types of masks, namely motion information temporal mask, road information spatial mask, and interaction relationship mask, are designed for self-supervised pre-training tasks on the model to enhance the model's ability to extract general scene features. Finally, pretraining weights are used as initialization parameters for supervised learning fine-tuning in downstream tasks. Experimental results on the Argoverse2 Motion Forecasting dataset show that the model can effectively reconstruct traffic scenes in pretraining tasks. The introduction of self-supervised pretraining improves prediction accuracy and data utilization efficiency. Moreover, it exhibits universality for different prediction tasks, achieving a 3.3% and 3.7% improvement in the $minFDE_6$ for single-agent and multi-agent trajectory prediction tasks, respectively.

Keywords: autonomous driving; trajectory prediction; self-supervised pretraining

前言

近年来,自动驾驶车辆是现代汽车智能化的重要发展方向。自动驾驶汽车不仅需要实时感知出周围场景信息,还需要理解交通场景信息并预测出未来一段时间内周围交通场景的变化趋势,准确的预测是自动驾驶车辆进行安全合理规划决策的基础。轨迹预测是根据已知交通场景中交通参与者历史轨迹与道路场景信息预测周围动态障碍物未来一段时

* 国家自然科学基金(52172382)、中央高校基本科研业务费项目(DUT22JC09)和辽宁省科学技术计划项目(2022JH1/10400030)资助。
原稿收到日期为 2024 年 01 月 23 日,修改稿收到日期为 2024 年 03 月 02 日。
通信作者:连静,副教授,博士,E-mail:lianjing@dlut.edu.cn。

间内的轨迹。

轨迹预测的研究方法可以分为两类:基于物理模型和基于学习的预测方法。基于物理模型的方法包括运动学模型、可达性分析[1]等,只能预测未来短时域内的运动轨迹,适用于辅助驾驶系统中的防撞预警。基于学习的轨迹预测方法能够从大量数据中识别出最有可能的运动模式;浅层机器学习方法依赖于人工特征工程,包括基于动态贝叶斯网络的运动估计[2]、基于隐马尔科夫模型的运动模式分类[3]等;基于深度学习的方法凭借强大的特征表示能力和性能成为主流,VectorNet[4]使用矢量化的场景表示形式及较小归纳偏置的轻量化网络结构开创了先例;MultiPath++[5]使用上下文门控机制及动态可学习轨迹锚点,兼顾预测性能与实时性;THOMAS[6]通过图神经网络与考虑碰撞可能性的热力图输出表示取得较好的预测性能;GANet[7]将轨迹预测任务拆分为目标区域预测与轨迹预测的两阶段方法来分别建模意图不确定与控制不确定性;GSA[8]改进Transformer[9]结构和目标引导取得优异的预测性能。

目前基于深度学习的轨迹预测算法大多仅使用监督学习[4-8],监督学习受数据噪声、标签形式的影响,存在数据利用效率低、模型精度有限的问题[10]。为了实现更精确的预测,模型规模越来越大,同时也需要更大规模数据,更大规模的模型与数据导致算力消耗剧增,已经成为目前成为亟待解决的问题。与传统的监督学习不同,自监督预训练不需要人工标注的标签,而是通过设计一个自动生成的标签来进行训练,使模型在自监督预训练阶段学习到一种通用的特征表达并用于下游任务。

自监督预训练方法主要分为两类:一类是基于对比学习(contrastive learning),MoCo[11]通过使图像正样本和负样本的差异最大化的对比学习进行预训练,使模型学习到图像通用的特征表示,并在下游图像分类、目标检测和图像生成任务中超过了监督学习方法的性能;另外一类基于掩码重建(masked modeling),BERT[12]通过随机掩码输入文本后重建和下文判断的预训练任务,在机器翻译、文本理解等多个下游任务中取得了显著的突破;GPT[13]通过使用自回归的预训练方式,在文本生成方面表现出色。目前,一些工作已将自监督预训练应用至轨迹预测任务中,PreTraM[14]通过构建历史轨迹与周围场景之间的对比学习预训练任务,实现在70%训练数据上达到相似预测精度;Traj-MAE[15]、RMP[16]等方法已证明通过掩码重建的自监督预训练能够提升已有预测模型精度;SEPT[17]通过对模型不同结构分别使用不同类型掩码重建,并成为Argoverse2数据集[18]的SOTA算法;Forecast-MAE[19]仅对Transformer结构掩码重建出历史轨迹和道路结构,提升了模型预测精度。基于以往工作中对自监督预训练引入轨迹预测的尝试,本文重点对掩码重建所适用的网络结构、掩码类型进行研究。

本文提出一种基于Transformer的轨迹预测模型及自监督预训练策略,如图1所示。预测模型由时序编码器、空间编码器、特征融合模块、重建头和轨迹预测头构成,能够提取场景内的时序、空间信息并建模交互关系。训练策略分为预训练与微调阶段。预训练阶段通过运动信息时序掩码、道路信息空间掩码、交互关系掩码3种掩码重建任务,提升模型对通用特征的提取能力;在微调阶段,以预训练阶段的模型为初始化参数,在下游任务中进行轨迹预测的监督学习。

1 场景表示及预测模型

1.1 场景表示

对于一个交通场景,交通参与者的历史运动状态可表示为$Agent_{\text{history}} \in \mathrm{R}^{A\times T_{\text{h}}\times D_{\text{a}}}$,交通参与者的未来运动状态信息为$Agent_{\text{future}} \in \mathrm{R}^{A\times T_{\text{f}}\times D_{\text{a}}}$,其中$A$为场景内交通参与者数量,包括预测目标、主车、周围其他交通参与者,T_{h}为历史观测时长,T_{f}为预测未来时长,D_{a}为交通参与者的运动状态特征;预测目标的未来运动轨迹坐标被表示为$Traj_{\text{gt}} \in \mathrm{R}^{A_{\text{target}}\times T_{\text{f}}\times 2}$,其中$A_{\text{target}}$表示需要被预测的目标数量,轨迹特征为俯瞰视角下交通参与者的二维坐标,使用向量化的道路表示。将车道中心线、车道边界等距离采样为多个道路点,并将相邻的N个道路点合并为车道段$Polyline \in \mathrm{R}^{N\times D_{\text{r}}}$。场景中的所有车道段构成道路信息$Road \in \mathrm{R}^{M\times N\times D_{\text{r}}}$,$M$为场景内车道段的数量,$D_{\text{r}}$为道路点特征,包括每个道路点的位置、类型信息。

对于交通参与者的时序运动状态,将运动状态拆分为当前时刻的绝对位置与时序每帧之间的相对运动状态:绝对位置包括平面二维坐标x、y和朝向角yaw,相对运动状态包括每一帧相对上一帧的位移d、速度变化量Δv、朝向角变化量Δyaw,能够保证运动特征的平移不变性,而通过Transformer的位置编码实现交通参与者在场景内位置的建模;类似

地,车道特征也被拆分为车道段几何中心点的绝对位置与其他点相对位置,以实现车道信息的平移不变形,使用车道段几何中心点位置坐标进行位置编码。

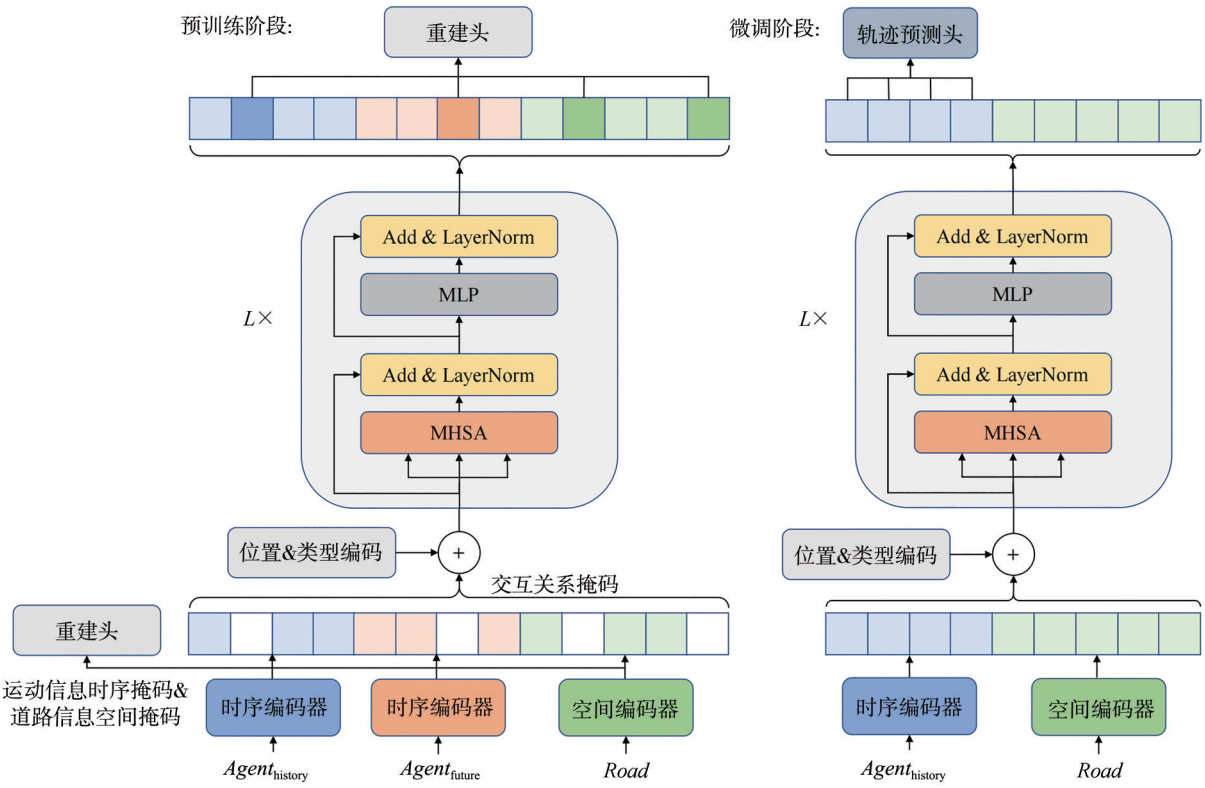


图1　方法总体示意图

1.2　预测模型结构

模型结构包括时序编码器、空间编码器、特征融合模块、重建头、轨迹预测头。在预训练和微调阶段模型架构略有差异,可以兼容不同序列长度的令牌。

对于时序编码器,虽然循环神经网络、Transformer可以处理时序数据,但存在实时性较差的问题,使用由一维卷积(1d CNN)构成的多尺度特征融合结构对提取时序特征,在克服RNN无法并行计算的缺点的同时,保持轻量化的结构。时序编码器的结构如图2所示,由3层一维卷积堆叠而成,每层使用尺寸为3的卷积核进行下采样,对每层输出使用线性插值上采样后拼接,最终由多层感知机(MLP)和层归一化(layer normalization)输出时序编码特征,通过残差连接来提升模型对不同尺度时序特征的提取能力。

在预训练阶段,两个相同结构的时序编码器分别编码交通参与者的历史运动状态信息和未来运动状态信息,从中得到交通参与者的历史运动嵌入$Embed_{\text{history}} \in \mathrm{R}^{A\times D}$和未来运动嵌入$Embed_{\text{future}} \in \mathrm{R}^{A\times D}$,$D$为模型特征通道数;在微调阶段,仅使用一个时序编码器对历史运动状态特征进行编码,得到每个交通参与者的历史运动嵌入$Embed_{\text{future}} \in \mathrm{R}^{A\times D}$。

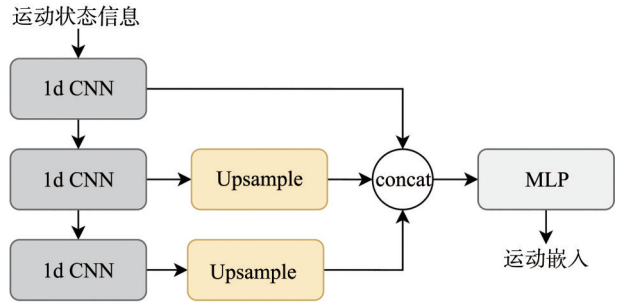


图2　时序编码器结构图

对于空间编码器,鉴于道路信息中包含数量众多的车道段,需要一种轻量化且能够兼容不同数量车道段的空间编码结构,使用由MLP与最大池化层组成的模块,结构如图3所示,通过最大池化提取全局道路特征,并通过残差连接后的拼接来融合全局特征和车道段特征,道路信息空间编码器后得到道路嵌入$Embed_{\text{road}} \in \mathrm{R}^{M\times D}$。

将上述不同类型嵌入拼接构成输入令牌序列,

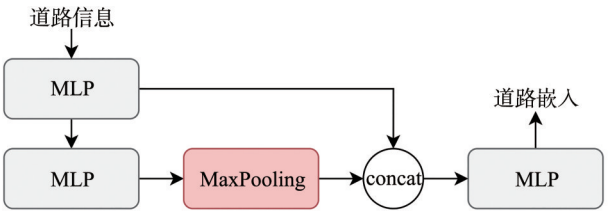


图3 空间编码器结构图

以表示场景全局信息。在特征融合模块中,通过对输入令牌序列进行特征融合,以建模交通参与者及车道间交互关系,并提取场景级别的特征。Transformer能够通过注意力机制提取序列中的全局特征信息,实现对不同特征的融合,克服了特征加权求和、拼接等传统的特征融合方法难以建模特征之间的复杂关系的缺点。特征融合模块使用标准的Transformer编码器,包括多头自注意力层、MLP、层归一化:在多头自注意力中,通过缩放点积注意力计算每个位置与其他位置的相关性权重,以动态分配注意力权重来融合序列信息。同时,通过残差连接和层归一化来缓解梯度消失,并通过堆叠L层Transformer编码器来增强模型的表示能力。

为了使Transformer识别输入序列的位置信息与类型信息,在输入序列中添加位置编码(positional encoding)和类型编码(type encoding)。位置信息及类型信息采用MLP嵌入实现动态编码,相较于正弦编码,MLP嵌入能够更灵活地建模不同位置信息;对于交通参与者的历史或未来运动状态信息,使用当前时刻的坐标和车头朝向作为位置信息进行位置编码;对于车道信息,对车道段几何中心点进行MLP嵌入作为其位置编码。为了区分不同嵌入所属类型,对不同类型的交通参与者和车道设置相应的可学习矩阵作为类型编码。

在预训练阶段,将交通参与者的历史运动状态嵌入、未来运动状态嵌入、车道嵌入拼接后作为输入令牌传入特征融合模块,并加入位置编码与类型编码,得到预训练阶段的输出令牌$T_{\text{Pretrain}} \in \mathrm{R}^{(2A+M)\times D}$。

在微调阶段,由于未来运动信息是监督学习的标签,不作为模型输入。将历史运动状态嵌入、车道嵌入拼接作为输入令牌序列,建模不同交通参与者、车道段之间的交互,得到微调阶段的输出令牌$T_{\text{Fintune}} \in \mathrm{R}^{(A+M)\times D}$。

预测头由MLP构成,对输出令牌中预测目标所属令牌$T_{\text{Finetune}}[A_{\text{target}}] \in \mathrm{R}^{A_{\text{target}}\times D}$进行解码,得到预测置信度$p \in \mathrm{R}^{A_{\text{target}}\times K}$与预测轨迹分布$Traj_{\text{pred}} \in \mathrm{R}^{A_{\text{target}}\times K\times T_f\times 5}$。为了描述交通参与者未来多种可能性的轨迹,对每个预测目标同时预测K条运动模态轨迹,并对预测轨迹使用高斯混合模型(Gaussian mixture model)表示其分布:

$$Traj_{\text{pred}} = \sum_{i=1}^{K}\sum_{t=1}^{T_f} N(\mu_x,\mu_y,\sigma_x,\sigma_y,\rho) \tag{1}$$

式中:$Traj_{\text{pred}}$为预测轨迹的分布;K为预测轨迹的运动模态数;T_f为未来轨迹帧数;μ_x、μ_y为预测坐标的均值;σ_x、σ_y为预测坐标的标准差;ρ为协方差系数;$N(\cdot)$为二维高斯分布。

2 自监督预训练策略

2.1 掩码重建任务

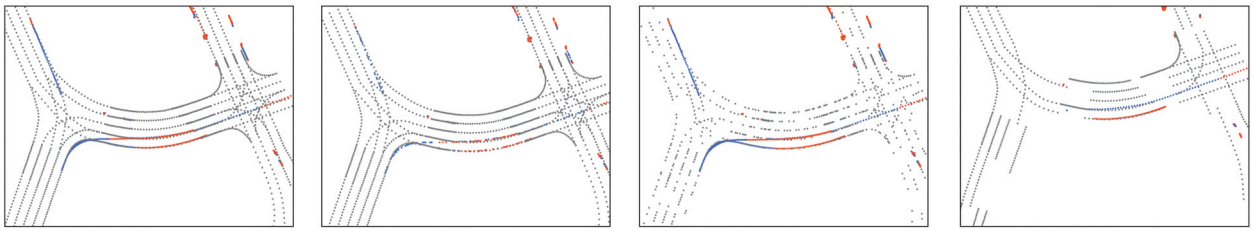
掩码重建是将输入样本的部分数据隐藏,并使用神经网络对被掩码部分进行重建,从而创建自监督预训练任务。BERT和MAE[20]已证明模型能够通过掩码重建任务学习到通用的特征表示来提升下游任务的性能。

与BERT中仅存在时间序列关系、MAE中仅存在空间关系不同,轨迹预测的数据中同时存在时序与空间关系。在本文掩码重建预训练任务中,设计了3种类型掩码:运动信息时序掩码、道路信息空间掩码、交互关系掩码。不同类型掩码可视化如图4所示,蓝色点表示交通参与者的历史轨迹、红色点表示交通参与者的未来轨迹,灰色点表示道路点。在模型结构中,时序编码器、空间编码器、特征融合模块分别与由MLP构成的重建头组成自编码器(autoencoder),以重建被掩码信息。

对交通参与者的历史及未来运动状态信息进行时序掩码,如图5所示,在运动状态信息输入时序编码器之前,随机对时间维度上的运动特征按照比例R_a进行掩码,即将该交通参与者在该时间步的所有运动特征置为零,并在有效位上标注为无效,之后历史与未来时序状态信息分别经过两个相同结构的时序编码器,得到每个交通参与者各自的运动嵌入,最后通过重建头输出被掩码轨迹点的坐标。

类似地,如图6所示,在输入空间编码器之前对车道信息在道路点维度上以比例R_b进行掩码,将车道段的被掩码道路点所有特征置为零,空间编码器与重建头构成自编码器,输出被掩码的道路点的坐标。

在特征融合模块中,如图1所示,每个输入令牌



(a) 原始场景　(b) 运动信息时序掩码　(c) 道路信息空间掩码　(d) 交互关系掩码

图4　不同类型掩码可视化

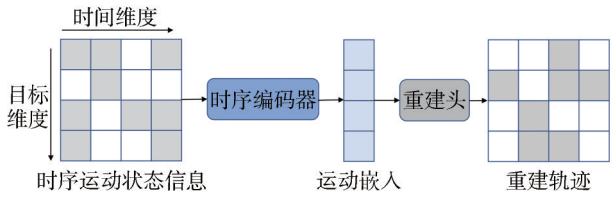


图5　时序掩码重建任务示意图

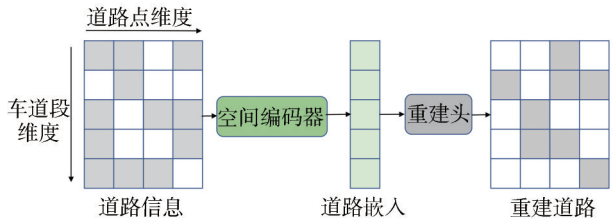


图6　空间掩码重建任务示意图

代表不同交通参与者的运动嵌入或不同车道段的车道嵌入,为了使模型建模不同交通参与者、不同车道段之间的交互关系,进行令牌尺度的随机掩码操作,掩码比例为R_c,将被掩码的嵌入置零,保留位置编码与类型编码,最后由重建头生成被掩码令牌嵌入前所对应轨迹或道路点坐标。

2.2　模型训练策略

本方法使用预训练-微调(pretrain-finetune)的训练策略。首先在训练集使用上述3类掩码重建任务对模型进行预训练,之后使用预训练权重对模型中的时序编码器、空间编码器和特征融合模块进行参数初始化,最后对模型在轨迹预测任务的训练集中端到端微调。

在预训练阶段,将所有被掩码部分重建点坐标与真值的L2距离作为预训练阶段损失函数:

$$Loss_{\mathrm{pt}}=\alpha\sum_{i=1}^{P}\left(Traj_{\mathrm{rc}}-Traj_{\mathrm{gt}}\right)^2+(1-\alpha)\sum_{i=1}^{Q}(Lane_{\mathrm{rc}}-Lane_{\mathrm{gt}})^2 \tag{2}$$

式中:α为损失调整系数;P为被掩码数据对应轨迹点数量;Q为被掩码数据对应道路点数量;$Traj_{\mathrm{rc}}$为重建轨迹点坐标;$Traj_{\mathrm{gt}}$为轨迹点真值坐标;$Lane_{\mathrm{rc}}$为重建道路点坐标;$Lane_{\mathrm{gt}}$为道路点真值坐标。

在微调阶段,为了鼓励预测模型预测多运动模态的预测轨迹,使用了赢者通吃策略,仅对与真值最为接近的预测轨迹计算回归损失,并对不同运动模态轨迹的置信度进行硬分配,通过对预测高斯混合分布使用最大期望化算法进行优化,回归损失使用预测值与真值的L2距离,而分类损失使用logsoftmax损失,损失函数的计算公式:

$$Loss_{\mathrm{ft}}=-\sum_{k=1}^{K}O(1)\left[\log\operatorname{softmax}\left(p^{*}|p\right)+\sum_{t=1}^{T_{\mathrm{f}}}\log N\left(Traj_{\mathrm{gt}}|Traj_{\mathrm{pred}}\right)\right] \tag{3}$$

式中:K为预测的运动模态数;T_{f}为未来轨迹帧数;$O(1)$表示仅当该运动模态轨迹距真值最近时为1其余为0;p为预测置信度;p^{*}为运动模态分类真值独热编码;$Traj_{\mathrm{gt}}$为真值轨迹;$N(\cdot)$为二维高斯分布;K为预测轨迹的运动模态数。

3　实验结果及分析

3.1　实验设置

实验使用Argoverse2 Motion Forecasting数据集,数据集由在迈阿密等6座城市中采集,覆盖城区及高速路段各类交通场景,训练集包含199 908个样本、验证集包含24 988个样本,每个场景按照10 Hz的频率采集。根据过去5 s的历史场景信息,预测目标在未来6 s的轨迹坐标,预测目标包含车辆、行人、骑行者等交通参与者,涵盖了单目标轨迹预测和多目标轨迹预测两种任务。Argoverse2数据集具有足够的数据量和丰富的场景,可以作为实验评估的基准系统,并与其他轨迹预测模型进行合理比较。

实验沿用Argoverse2中评价指标,使用预测轨

迹与未来轨迹真值之间的误差来衡量预测性能,包括最小平均位移误差(minADE)、最小终点误差位移(minFDE)、错失率(MR),单运动模态仅考虑置信度最高单条预测轨迹,多运动模态考虑置信度最高的6条轨迹。

基于PyTorch深度学习框架实现模型搭建,并在GeForce RTX 3060 GPU硬件平台上实验。在模型结构上,设置特征通道数D为128,Transformer中自注意力的头数为8,特征融合模块中Transformer层数L设置为4。对于优化方法,预训练及微调阶段的设置保持一致,均使用Adam优化器,初始学习率为0.001,Cosine学习率调度器,训练轮次为60个,并设置0.000 1的权重衰退来抑制模型过拟合;在预训练中,设置3类掩码比例R_a、R_b、R_c为0.5;对未进行预训练的Scratch模型使用Xavier[21]随机初始化。

3.2 轨迹预测性能

为了验证本方法的预测性能,在单目标轨迹预测和多目标轨迹预测两个任务中进行了对比实验。对比对象包括引入预训练的模型Pretrained、未进行预训练的模型Scratch以及目前主流预测算法。

使用Argoverse2数据集进行单目标轨迹预测任务,其中每个交通场景仅包含一个预测目标。本方法模型及部分主流预测算法在测试集中预测误差及排行榜名次如表1所示。

表1 单目标轨迹预测性能指标

模型	minADE_1	minADE_6	minFDE_1	minFDE_6	MR_1	MR_6	排行榜名次
MultiPath++[5]	2.14	0.86	5.34	1.71	0.66	0.25	59
THOMAS[6]	1.96	0.88	4.62	1.51	0.64	0.20	45
Forecast-MAE[19]	1.76	0.80	4.38	1.41	0.61	0.18	31
GANet[7]	1.78	0.73	4.48	**1.35**	0.61	**0.17**	36
本方法 Scratch	1.791	0.732	4.484	1.453	0.613	0.186	37
本方法 Pretrained	**1.733**	**0.711**	**4.332**	1.403	**0.590**	0.173	**26**

注:Multipath++算法的性能指标由本文作者对该开源算法在Argoverse2数据集训练测试得到;其余对比算法的性能指标优先使用文中的测试结果,若文中未提供全部指标测试结果,使用EvalAI平台中Argoverse2数据集的排行榜结果;本文中排行榜名次以minFDE_1指标进行排序,Argoverse2数据集排行榜链接:https://eval.ai/web/challenges/challenge-page/1719。

引入自监督预训练后,模型表现出更小的预测误差,其中单运动模态指标minADE_1减少3.2%、minFDE_1减少3.4%、MR_1减少3.8%;多运动模态指标minADE_6减少2.6%,minFDE_6减少3.3%,MR_6减少7.4%;此外与MultiPath++、THOMAS等预测算法相比,本方法在大部分指标上都取得了领先的结果,在Argoverse2数据集中排名第26名。

对比训练过程中的是否引入自监督预训练的训练损失曲线,如图7所示。Pretrained模型在首个轮次的损失值明显低于Scratch模型,并且在训练结束时,Pretrained模型收敛下界低于Scratch模型;在验证集损失值中也观察到类似的现象,但是Pretrained模型的minFDE_6指标在训练中期发生较大波动,可能是此时学习率过大或分类损失与回归损失优化不平衡所致,在训练末期学习率减小时,Pretrained模型的收敛下界同样低于Scratch模型,因此使用本文提出的预训练任务有助于模型快速收敛并提升预测精度。

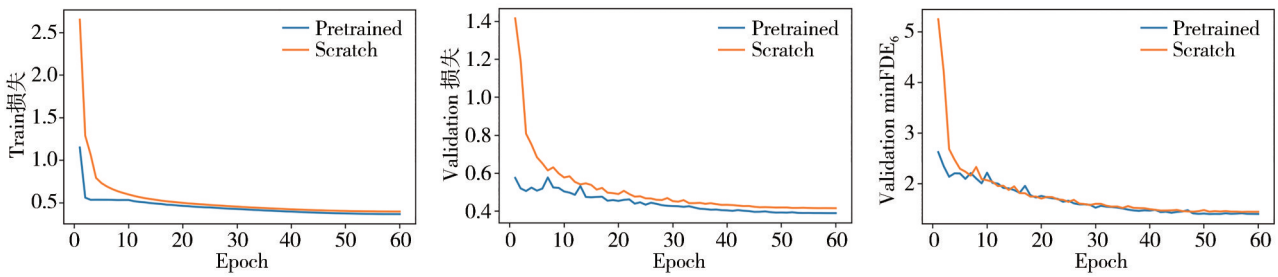


图7 损失函数及minADE_6指标曲线

本方法的模型可以兼容多目标轨迹预测任务,与单目标轨迹预测使用相同的预训练权重进行微调,而无须重新预训练。在Argoverse2数据集中进行多目标轨迹预测实验,结果如表2所示,与最新的

多目标轨迹预测算法 FJMP[22]、FFINet[23]相比,本方法在多运动模态预测精度表现领先,单运动模态预测性能也具有竞争力。此外,使用自监督预训练的 Pretrained 模型在各指标上都优于未进行预训练的 Scratch 模型,其中 minFDE_6 指标提升 3.7%,说明自监督预训练对不同的下游任务都有精度提升,并具有一定的通用性。

3.3　自监督预训练的消融实验

为了验证自监督预训练中掩码重建的效果,对不同类型的掩码重建任务进行消融实验,通过对比使用不同掩码类型预训练权重微调后的预测精度结果来验证各类掩码重建任务的有效性。

表 2　多目标轨迹预测性能指标

模型	minADE_1	minADE_6	minFDE_1	minFDE_6	MR_1	MR_6
FJMP[22]	1.52	0.81	4.00	1.89		0.23
FFINet[23]	**1.24**	0.77	**3.18**	1.77		0.24
本方法 Scratch	1.401	0.734	3.668	1.676	0.334	0.200
本方法 Pretrained	1.368	**0.708**	3.576	**1.615**	**0.333**	**0.196**

如表 3 所示,首先只使用单种类型掩码重建任务,相较未进行预训练的 Scratch 模型,所有类型的掩码重建任务都带来了精度的提升,其中交互掩码的提升幅度最大;之后对多种掩码混合使用的效果进行实验,发现同时使用时序掩码、空间掩码和交互掩码时模型表现最优,说明 3 种掩码在叠加使用时没有发生负面干扰,均对模型性能提升有正面效果;此外,时序掩码及空间掩码中的性能提升说明掩码重建对本文的时序编码器及空间编码器结构同样有效,证明了掩码重建任务对 Transformer 以外部分结构的有效性。

表 3　不同类型掩码重建任务的消融实验

时序掩码	空间掩码	交互掩码	minADE_6	minFDE_6	MR_6
√			0.726	1.396	0.190
	√		0.735	1.402	0.218
		√	0.709	1.367	0.177
√	√		0.723	1.389	0.191
√		√	0.707	1.355	0.179
√	√	√	**0.703**	**1.344**	**0.174**
Scratch			0.759	1.418	0.221

3.4　数据效率实验

为了验证所提出的自监督预训练对于数据效率的影响,从 Argoverse2 训练集中随机抽取一定比例的数据样本进行预训练与微调,并与全量训练集训练且未经预训练的 Scratch 模型在验证集上对比预测性能,结果如表 4 所示。

当训练样本减少时,模型的预测性能也随之下降;当抽取比例为 0.6 时,预测性能仍优于使用全量训练集数据的 Scratch 模型,引入预训练后,仅使用 60% 的数据便可达到与 Scratch 模型相当的预测精度,说明所提出自监督预训练策略提高了数据的利用效率。

表 4　数据效率实验

抽取比例	minADE_6	minFDE_6	MR_6
0.8	0.711	1.352	0.175
0.7	0.739	1.388	0.186
0.6	0.753	1.411	0.214
0.5	0.783	1.496	0.227
Scratch	0.759	1.418	0.221

3.5　模型规模及实时性

对本方法及部分开源预测算法的模型规模和实时性进行测试,使用模型中可学习参数量和计算量来衡量模型规模,模型前向推理耗时衡量模型的实时性,测试结果如表 5 所示。

表 5　模型参数量及推理耗时

模型	参数量/Million	计算量/GFlops	推理耗时/ms
MultiPath++	21.1	3.62	16.3
Forecast-MAE	1.92	0.949	5.03
本方法	**1.25**	**0.692**	**3.37**

是否进行预训练不影响模型参数量和推理耗时,所以本方法的 Scratch 模型与 Pretrained 模型在模型规模及实时性上表现一致。本方法模型可学习参数量仅为 125 万个,计算量为 0.692 GFlops,在 GeForce RTX 3060 GPU 硬件平台上进行单目标预测时,模型推理平均耗时可以达到 3.37 ms,明显优于 MultiPath++、Forecast-MAE 开源算法,说明本方法模型是具备优异实时性的轻量化模型。

3.6　结果可视化

对预训练掩码重建任务、单目标轨迹预测、多目

标轨迹预测在 Argoverse2 验证集部分典型场景进行可视化,并对比是否使用自监督预训练的预测结果。

掩码重建任务的可视化结果如图 8 所示,其中左图为未掩码的真值场景,右图为掩码后重建的预测场景,灰色点为道路采样点的真值、蓝色点为交通参与者轨迹真值,橙色点为预训练任务所生成的重建点;模型能够重建出被掩码的车道和轨迹的位置及形状,同时符合轨迹点之间的插值关系和车道的连接关系,说明本文提出的自监督预训练任务能够在一定程度上使模型理解交通场景。

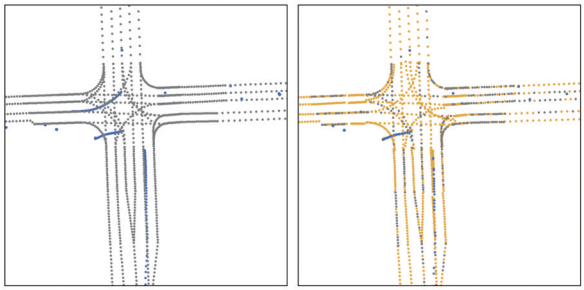


图 8 掩码重建任务

单目标轨迹预测结果可视化如图 9 和图 10 所示,多目标轨迹预测可视化如图 11 所示,橙色方框或圆点代表预测目标,蓝色方框或圆点代表周围其他交通参与者,蓝色线条表示历史观测轨迹,橙色线条表示预测轨迹,红色线条表示未来轨迹真值。

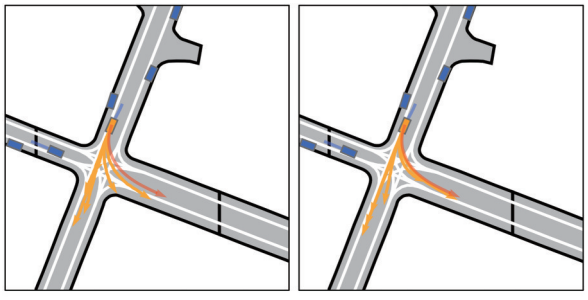


图 9 车辆轨迹预测

在图 9 中,对城区十字路口内车辆目标进行预测,两个模型均预测出车辆的直行和左转轨迹,但 Scratch 模型的左转轨迹转向曲率过小,有一条轨迹即将超出道路边界,Pretrained 模型能够根据道路结构预测出合适的转向曲率轨迹,并且还包含不同速率下的转向和直行的轨迹,从而兼顾多运动模态与准确性。

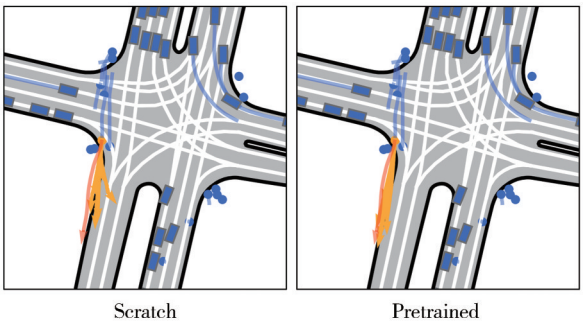


图 10 骑行者轨迹预测

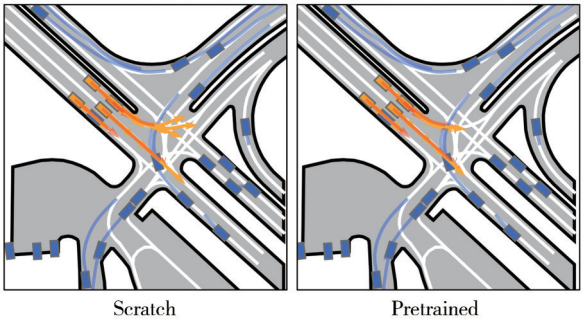


图 11 多目标预测

在图 10 中,对城区路口的骑行者穿行马路后的轨迹进行预测,由于骑行者样本在数据集中占比较少,预测骑行者的行为具有挑战性。对于骑行者,Scratch 模型预测其闯入沿道路中央的轨迹,并没有准确预测出骑行者穿过马路后进行加速的行为,而 Pretrained 模型能够准确预测出骑行者沿道路边界加速直行的轨迹。

在图 11 中,对多目标预测结果进行可视化,在复杂路口下车辆起步的行为进行预测,模型能够预测出相同车道的车辆行为具有一定相似性,模型表现出捕获不同预测目标之间的相关系的能力,这对于多目标预测是至关重要的;此外,经过预训练的 Pretrained 模型预测更加准确。

4 结论

本文提出了一种基于轨迹预测模型及通用的自监督预训练策略,搭建了基于 Transformer 的轻量化轨迹预测模型,设计了 3 类掩码重建任务,包括运动信息时序掩码、道路信息空间掩码、交互关系掩码,用于自监督预训练。Argoverse2 数据集中的实验表明,本文提出的模型与目前主流预测模型相比在精度和推理耗时上具备竞争力;本文提出的自监督预

训练策略能够提升模型预测精度和数据利用效率,并对单目标与多目标轨迹预测任务具有通用性。

本文所提出的自监督预训练策略仍有深入研究的空间,例如验证其在受遮挡目标的轨迹预测、基于学习的路径规划等任务中的有效性和通用性,在更大规模数据集上验证预训练数据量对性能的提升等。

参考文献

[1] KOSCHI M, ALTHOFF M. SPOT: a tool for set-based prediction of traffic participants[C]. IEEE Intelligent Vehicles Symposium, 2017.

[2] SCHREIER M, WILLERT V, ADAMY J. An integrated approach to maneuver-based trajectory prediction and criticality assessment in arbitrary road environments[J]. IEEE Transactions on Intelligent Transportation Systems, 2016.

[3] WEI Y, ZHEN L, WANG C, et al. Lane-change prediction method for adaptive cruise control system with hidden Markov model[J]. Advances in Mechanical Engineering, 2018.

[4] GAO J, SUN C, ZHAO H, et al. VectorNet: encoding HD maps and agent dynamics from vectorized representation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[5] VARADARAJAN B, HEFNY A, SRIVASTAVA A, et al. MultiPath++: efficient information fusion and trajectory aggregation for behavior prediction[C]. International Conference on Robotics and Automation (ICRA), 2022.

[6] GILLES T, SABATINI S, TSISHKOU D, et al. THOMAS: trajectory heatmap output with learned multi-agent sampling[C]. International Conference on Learning Representations (ICLR), 2022.

[7] WANG M, YU C, WANG M, et al. GANet: goal area network for motion forecasting[C]. International Conference on Robotics and Automation (ICRA), 2023.

[8] 连静,李硕贤,刘一荻,等. 基于车道目标引导的车辆轨迹预测[J]. 汽车工程, 2023, 45(8): 1353-1361.
LIAN J, LI S, LIU Y, et al. Goal supervised attention network for vehicle trajectory prediction[J]. Automotive Engineering, 2023, 45(8): 1353-1361.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Neural Information Processing Systems, 2017.

[10] BALESTRIERO R, IBRAHIM M, SOBAL V, et al. A cookbook of self-supervised learning [J]. arXiv preprint: 2304.12210, 2023.

[11] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020.

[12] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]. Proceedings of the 2019 Conference of the North, Minneapolis, Minnesota, 2019.

[13] FLORIDI L, CHIRIATTI M. GPT-3: its nature, scope, limits, and consequences[J]. Minds and Machines, 2020: 681-694.

[14] XU C, LI T, TANG C, et al. PreTraM: self-supervised pre-training via connecting trajectory and map[C]. European Conference on Computer Vision (ECCV), 2022.

[15] CHEN H, WANG J, SHAO K, et al. Traj-MAE: masked autoencoders for trajectory prediction[C]. International Conference on Computer Vision (ICCV), 2023.

[16] YANG Y, ZHANG Q, GILLES T, et al. RMP: a random mask pretrain framework for motion prediction[J]. arXiv preprint: 2309.08989, 2023.

[17] LAN Z, JIANG Y, MU Y, et al. SEPT: towards efficient scene representation learning for motion prediction[J]. arXiv preprint: 2309.15289, 2023.

[18] WILSON B, QI W, AGARWAL T, et al. Argoverse 2: next generation datasets for self-driving perception and forecasting[J]. Neural Information Processing Systems, 2021.

[19] JIE C, XIAODONG M, MING L. Forecast-MAE: self-supervised pre-training for motion forecasting with masked autoencoders[C]. International Conference on Computer Vision (ICCV), 2023.

[20] HE K, CHEN X, XIE S, et al. Masked autoencoders are scalable vision learners[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[21] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[C]. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010.

[22] ROWE L, ETHIER M, DYKHNE H, et al. FJMP: factorized joint multi-agent motion prediction over learned directed acyclic interaction graphs[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[23] KANG M, WANG S, ZHOU S, et al. FFINet: future feedback interaction network for motion forecasting[J]. IEEE Transactions on Intelligent Transportation Systems, 2023.