SIMPL: A Simple and Efficient Multi-agent Motion Prediction Baseline for Autonomous Driving

Lu Zhang[1], Peiliang Li[2], Sikang Liu[2], and Shaojie Shen[1]

摘要：

本文介绍了 SIMPL：一种简单且高效的自动驾驶多智能体运动预测基线。与具有高精度但重复计算的传统 agent-centric 的方法和具有折中精度和泛化能力的 scene-centric 的方法不同，SIMPL 为所有相关的交通参与者提供实时、准确的运动预测。为了提高准确性和推理速度，本文提出了一种紧凑且高效的 global feature fusion module，该模块以对称的方式执行定向消息传递，使网络能够在单个前馈过程中预测所有道路使用者的未来运动，并且降低由视点变换所造成的精度损失。此外，本文还在轨迹解码中 using Bernstein basis polynomials 来研究连续轨迹参数化，其允许在任何期望的时间点评估状态及其高阶导数，这对下游规划任务很有价值。作为一个强大的基线，SIMPL 相比于其它最先进的方法在 Argoverse 1 &2 运动预测基准上展现出极具竞争力的性能。此外，SIMPL 的轻量化设计和低推理延迟使其具有高度的可扩展性，并且有望用于现实世界的车载部署.

创新点的提出：

1. 使用什么的 scene-represention 保持编码方式的高效性且对视角变化的鲁棒性？

对于基于学习的运动预测，最重要的主题之一是 scene-represention。早期的方法通常将周围场景表示为多通道鸟瞰图像 [1]－[4]。相比之下，最近的研究越来越多地采用矢量化场景表示[5]-[13]，其中使用带有地理坐标的点集或多段线(polylines) 来标注位置和几何形状，从而提高保真度并扩大感受野。然而，对于光栅化和矢量化表示，都存在一个关键问题：我们应该如何为所有这些元素选择合适的参考系？一种直接的方法是描述共享坐标系（scene-centric）内的所有实例，例如以自动驾驶车辆为中心的坐标系，并直接使用坐标作为输入特征。这使我们能够在一次前馈传递中对多个目标代理进行预测 [8, 14]。然而，使用全局坐标作为输入，通常会在单个前馈传递中对多个目标代理进行预测 [8, 14]。然而，使用全局坐标作为输入（通常会在很大范围内变化）将大大加剧任务的固有复杂性，导致网络性能下降和对新场景的适应性有限。为了提高准确性和鲁棒性，一种常见的解决方案是根据目标代理的当前状态对场景上下文进行归一化处理 [5, 7, 10]-[13]（agent-centric）。这意味着必须对每个目标代理重复执行归一化过程和特征编码，从而获得更好的性能，但代价是冗余计算。因此，有必要探索一种能够有效地编码多个目标的特征，同时保持对视角(perspective) 变化的鲁棒性的方法。

作者提出一种 Instance-centric 方法，与传统的 agent-centric 和 scene-centric scene represention 不同，其为每个 agent 轨迹 polyline 和 map polyline 建立属于自己的坐标系（agent-centric 和 scene-centric represention 只为 agent polyline 建立坐标系，map polyline 归一化到这些坐标系中，如图 1 所示），然后将 agent 轨迹 polyline 和 map polyline 分别进行编码得到 agent token 和 map token。但是，这么多 polylines 之间的位置关系怎么建立？作者提出 Relative Positional Encoding（RPE），维度（N,N,5），表示 N 个 polylines 之间的相对位置（N 为 agent 轨迹 polylines 和 map polylines 的总数量）编码。作者继续将 agent token 和 map token 沿着拼接后得到 scene tokens（N,128），将其沿着第零维度和第一维度 expand&repeat,得到 source tokens(N,N,128),target tokens(N,N,128),然后将 scene token、source tokens、arget tokens 三个矩阵拼接融合(symmetric fusion Transformer (SFT))，如 Figure 2 所示。在实际 debug 中，batchsize 设置为 32 时，N 的数值在 1000 左右，三个矩阵拼接后所占的内存量很大，这导致了模型训练速度的降低，同时通过简单的矩阵拼接融合来建立整个场景之间的位置关系，使模型不能充分捕捉 agent 和 map 之间的交互，导致整个模型在长序列 Argoverse 2 数据集上 表现不佳。

Fig1 Instance-centric、Scene-centric、Agent-centric 示意图

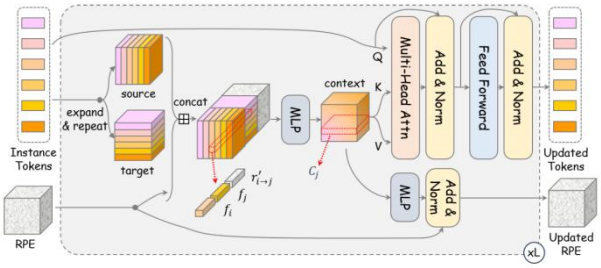

Figure 2.symmetric fusion Transformer (SFT)

2. 如何评估未来轨迹的航向、速度和其他高阶导数？

对于运动预测的下游模块，例如决策和运动规划，不仅需要考虑未来位置，还需要考虑航向、速度和其他高阶导数。例如，周围车辆的预测航向在塑造未来时空占用方面发挥着关键作用，这是确保安全和稳健的运动规划的关键因素[15, 16]。此外，在不遵守物理约束的情况下独立预测高阶量可能会导致预测结果不一致[17, 18]。例如，尽管速度为零，但它可能会产生位置位移，从而导致规划模块混乱。

本文引入了一种基于 Bernstein basis polynomial（也称为 Bezier curve）的预测轨迹的新颖参数化方法。这种连续表示确保了平滑性，并能够在任何给定时间点轻松评估精确状态及其高阶导数。本文的实证研究表明，与估计 monomial basis polynomials 的系数相比，学习预测 Bezier curves 的控制点更加有效且数值稳定。

与传统完全基于神经网络的解码器不同，该解码器通过将引入将不可训练的参数化曲线和可训练的神经网络结合，提高推理速度，但是参数化曲线本质上是通过数学公式回归出未来轨迹，其解码轨迹缺少多模态性，导致精度受限，这里可以采用 DETER-LIKE 形式的解码器提高其精度。

实验结果：

TABLE I: Results on the test split of Argoverse 1 motion forecasting dataset. The upper and lower groups are the results of single model and ensemble methods. The best result is in **bold** while the second best result is underlined. b-minFDE$_6$ is the official ranking metric. ♯ denotes the model size is from the non-official implementation

Method	minADE$_6$	minFDE$_6$	MR$_6$	b-minFDE$_6$	#Param
LaneGCN [6]	0.870	1.362	16.2	2.053	3.7M
mmTrans [7]	0.844	1.338	15.4	2.033	2.6M
SceneTrans [8]	0.803	1.232	12.6	1.887	15.3M
HiVT [9]	**0.774**	**1.169**	12.7	1.842	2.5M
MacFormer [12]	0.819	1.216	**12.1**	1.827	2.4M
SIMPL (w/o ens)	0.793	1.179	12.3	**1.809**	**1.8M**
MultiPath++ [30]	0.790	1.214	13.2	1.793	21.1M♯
MacFormer [12]	0.812	1.214	12.7	1.767	2.4M
HeteroGCN [13]	0.789	1.160	11.7	1.751	-
Wayformer [36]	**0.768**	1.162	11.9	**1.741**	11.2M♯
SIMPL (w/ ens)	0.769	**1.154**	**11.6**	1.746	**1.8M**

TABLE II: Results on the Argoverse 2 test split for methods based on symmetric scene modeling. The results are from single models (w/o ensemble). The best and the second-best results are in **bold** and underlined, respectively. b-minFDE$_6$ is the official ranking metric.

Method	minADE$_6$	minFDE$_6$	MR$_6$	b-minFDE$_6$	#Param
HDGT [19]	0.84	1.60	21.0	2.24	12.1M
GoRela [20]	0.76	1.48	22.0	2.01	-
QCNet [21]	**0.65**	**1.29**	**16.0**	**1.91**	7.3M
SIMPL (w/o ens)	0.72	1.43	19.2	2.05	**1.9M**

TABLE III: Ablative study of the feature fusion module design on the Argoverse 1 validation split.

Model	Emb. Size	# Layers	RPE Upd.	minFDE$_6$	MR$_6$	b-minFDE$_6$
$\mathcal{M}1$	64	2	✗	1.237	12.8	1.848
$\mathcal{M}2$	64	4	✗	1.037	9.5	1.658
$\mathcal{M}3$	128	4	✗	0.993	9.0	1.607
$\mathcal{M}4$	128	4	✓	0.947	**8.1**	1.559
$\mathcal{M}5$	128	6	✓	**0.944**	8.4	**1.558**

TABLE IV: Ablative study of the trajectory parameterization methods and the yaw angle loss on the Argoverse 2 validation set.

Parameterization	Yaw loss	minADE$_6$	minFDE$_6$	minAYE$_6$	minFYE$_6$
Raw coords	✗	**0.780**	**1.452**	0.134	0.151
Polynomial	✗	0.861	1.738	0.146	0.278
Bézier curve	✗	**0.780**	1.457	0.137	0.297
Bézier curve	✓	0.783	**1.452**	**0.055**	**0.076**

文章总结：

本篇文章提出了一种新的 scene represention--Instance centric,为每个 polylines 建立坐标系，并通过引入 RPE 建立整个场景 polylines 之间的位置关系，其通过扩展拼接后使用 symmetric fusion Transformer (SFT)融合 RPE 与 polylines，该方法省去大量的神经网络使参数量减少，但是需要占用过多的内存，同时简单的拼接这些矩阵使其在复杂数据集 Argoverse 2 上表现不佳；Bezier curve 作为解码器相对于 MLPs 来说，解码出的轨迹精度高，也提高了推理速度，不过，要想用这篇文章作为基线，可以通过置换编码器来提高预测精度。