城市驾驶环境下基于模型的模仿学习

摘要

一个准确的世界模型在改进运动规划方面具有巨大潜力。该论文提出了 MILE：一种基于模型的模仿学习方法，用于共同学习世界模型和自动驾驶策略。该论文利用 3D 几何作为归纳偏差，并直接从高分辨率专家演示视频中学习高度紧凑的潜在空间。该论文的模型能够预测多样且合理的状态和动作，可以被解释为鸟瞰图语义分割。此外，该论文展示它可以从完全在想象中预测的计划中执行复杂的驾驶动作。该论文的方法是第一个仅使用摄像头的方法，可以在城市驾驶环境中对静态场景、动态场景和自我行为进行建模。于此同时，我引入了具有特权模式的强化学习教练，其能提升自动驾驶在未知场景下的性能，它可以利用多种传感器信息来感知环境并输出结果。实验结果表明其在一定程度上有性能提升。

关键词：世界模型，模仿学习，强化学习教练

1 引言

人类是根据有限的感官来感受并理解世界，我们做出的决策大多情况下都依赖于我们的“世界模型”。当我们看到蜻蜓低飞时，我们或许会想到即将要下雨。基于该思想，该论文提出了一种基于模型的模仿学习方法，用于共同学习世界模型和自主驾驶的策略。该论文利用三维几何作为归纳偏置，直接从专家演示的高分辨率视频中学习高度紧凑的潜在空间。该模型可以预测各种合理的状态和行动，这些状态和行动可以被解释为鸟瞰式的语义分割。此外，它可以根据完全在想象中预测的计划执行复杂的驾驶动作。但是该论文使用模仿学习来模仿人类专家数据可能会使智能体在某些场景下表现不好。为此，我引入了具有特权模式的强化学习教练来提升其性能，它可以利用多种传感器信息来感知环境并输出结果。实验结果表明其在一定程度上有性能提升。

2 相关工作

现有的自动驾驶系统可大致分为三类：模块化设计，多任务框架，端到端。模块化设计方案中，每个独立的模块负责单独的子任务。这种方案具备简化研发团队分工，便于问题回溯，易于调试迭代等优点。但由于将不同任务解耦，各个模块相对于最终的驾驶规划目标存在信息损失问题，且多个模块间优化目标不一致，误差会在模块间传递。多任务框架中，不同任务使用同一个特征提取器，具备便于任务拓展、节省计算资源等优点。但不同任务之间

存在预测不一致、表征冲突的问题。针对上述问题，自动驾驶学术界和产业界将研究方向聚焦在了感知决策一体化上。通过端到端模型统一感知与决策两大体系，可以避免级连误差。

过去的研究缺乏对 3D 几何的利用，难以建立起静态和动态场景与自我行为之间的联系。文章提出的基于模型的模仿学习方法，使用 3D 几何作为归纳偏置，使得模型可以在城市驾驶环境中学习世界模型和驾驶策略，同时具有较好的泛化性能。文章首先使用一个生成模型表示城市驾驶环境，使用概率分布来表征自我行为、相对位置和驾驶操作等因素，并利用概率图模型来表示这些因素之间的关系；接着使用一个变分自动编码器来对后验分布进行估计，将观察数据映射到世界模型和自我行为中；最后，使用端到端的模仿学习来学习人类专家驾驶策略。于此同时，我引入了具有特权模式的强化学习教练，其能提升自动驾驶在未知场景下的性能，它可以利用多种传感器信息来感知环境并输出结果。

2.1 强化学习

最近的研究 [11] 展示了在使用真实尺寸的自动驾驶汽车时，采用深度强化学习（DDPG）的可能性。在这项工作中，系统首先经历了模拟训练，然后进行了实时训练，利用车载计算机，最终学会了沿车道行驶，并成功完成了一项在 250 米路段的真实测试。基于模型的深度强化学习算法还被用于直接从原始像素输入中学习模型和策略 [21] [20]。在 [4] 中，深度神经网络（DNNS）被用于在具有数百个时间步的模拟环境中生成预测。强化学习还适用于控制。在 [17] 中，经典的最优控制方法（如 LQR/iLQR 等）与强化学习方法进行了比较。经典的强化学习方法被用于在随机环境中执行最优控制，例如在线性状态下使用线性二次调节器（LQR），在非线性状态下使用迭代线性二次调节器（iLQR）。最近的一项研究 [14] 表明，采用参数的随机搜索的策略网络的性能与 LQR 相同。

2.2 模仿学习

早期的研究探讨了车辆驾驶的行为克隆（BC），参见 [15] [16]。该研究展示了示范性学习（LfD）代理，试图模仿专家的行为。行为克隆通常被实现为一种监督学习形式，这使得它难以适应新的不可见情境。在 [2] [3] 中，提出了一种用于自动驾驶车辆领域的架构，其中一个卷积神经网络（CNN）被端到端地训练，直接将来自单个前置摄像头的原始像素映射到转向命令。通过使用相对较小的来自人类或专家的训练集，系统学会在有或没有车道标线的局部路段和高速公路上驾驶车辆。网络学习可以成功地检测到道路的图像表示，无需显式训练。在 [13] 中，提出使用最大熵反向强化学习来学习通过人类驾驶员的专家演示实现舒适驾驶的轨迹优化。在 [19] 中，DQN 被用作反向强化学习中的一个精炼步骤，以导出奖励以学习类似人类的驾驶车道变更行为。

2.3 世界模型

基于模型的方法大多在强化学习环境中进行了探索，并被证明是非常成功的 [8,10,18]。尽管在完全离线强化学习方面已经取得了进展，但这些方法都假定获得奖励，并与环境进行在线互动 [22,23]。基于模型的模仿学习已经成为机器人操作 [6] 和 OpenAI Gym [12] 中强化学习的替代方案。尽管这些方法不需要获得奖励，但它们仍然需要与环境进行在线交互以获得良好的表现。从图像观测中学习世界模型的潜在动力学首先被引入到视频预测中 [1,5,7]。与

我们的方法最相似的是，[9,10] 另外建模了奖励函数，并在他们的世界模型中优化了策略。与之前的工作相反，我们的方法不假设访问奖励函数，而是直接从离线数据集中学习策略。此外，以前的方法操作简单的视觉输入，大多是 64×64 的大小。

3 本文方法

3.1 本文方法概述

文章提出的基于模型的模仿学习方法，使用 3D 几何作为归纳偏置，使得模型可以在城市驾驶环境中学习世界模型和驾驶策略，同时具有较好的泛化性能。文章首先使用一个生成模型表示城市驾驶环境，使用概率分布来表征自我行为、相对位置和驾驶操作等因素，并利用概率图模型来表示这些因素之间的关系；接着使用一个变分自动编码器来对后验分布进行估计，将观察数据映射到世界模型和自我行为中；最后，使用端到端的模仿学习来学习人类专家驾驶策略。于此同时，我引入了具有特权模式的强化学习教练，其能提升自动驾驶在未知场景下的性能，它可以利用多种传感器信息来感知环境并输出结果。

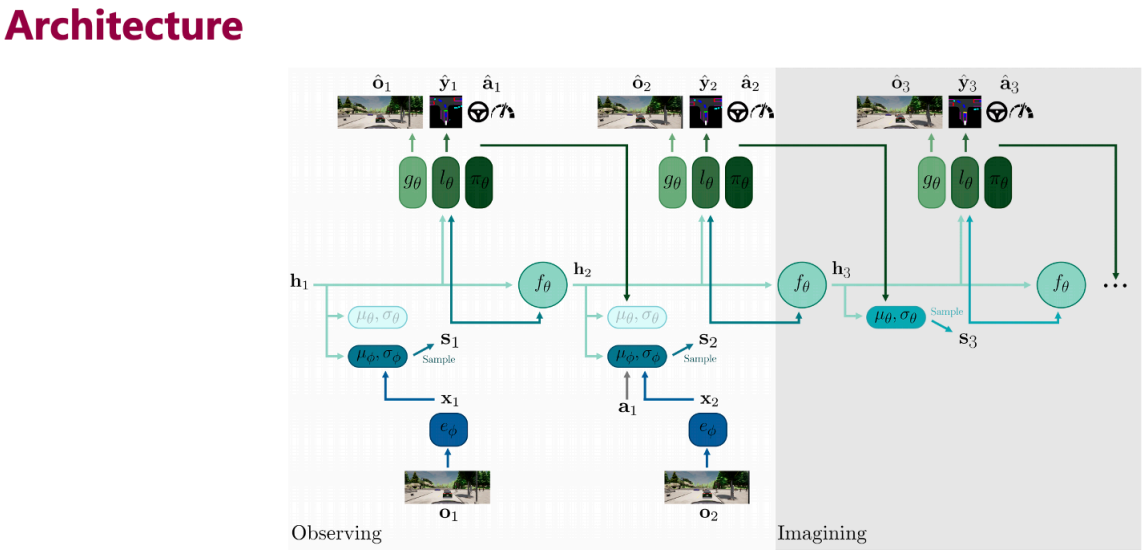

Figure 1: Architecture of MILE.

图 1. 架构

3.2 变量推断

$$p(\mathbf{o}_{1:T}, \mathbf{y}_{1:T}, \mathbf{a}_{1:T}, \mathbf{h}_{1:T}, \mathbf{s}_{1:T}) = \prod_{t=1}^{T} p(\mathbf{h}_t, \mathbf{s}_t | \mathbf{h}_{t-1}, \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) p(\mathbf{o}_t, \mathbf{y}_t, \mathbf{a}_t | \mathbf{h}_t, \mathbf{s}_t)$$

为了最大化观测数据的边际似然 $p(\mathbf{o}1:T, \mathbf{y}1:T, \mathbf{a}1:T)$，我们需要推断潜在变量 $\mathbf{s}1:T$。我们通过引入变分分布 $q_{\mathbf{H},\mathbf{S}}$ 来实现深度变分推断。

3.3 推断模型

推断网络，由参数 $\varnothing$ 参数化，建模 $q(\mathbf{s}_t|\mathbf{o}_{<t}, \mathbf{a}_{<t})$，其近似真实的后验分布 $p(\mathbf{s}_t|\mathbf{o}_{<t}, \mathbf{a}_{<t})$。该网络由两个元素组成：观测编码器 $\mathcal{E}_\varnothing$，用于嵌入输入图像、路径地图和车辆控制传感器数据到低维向量；后验网络 $(\mu_\varnothing, \sigma_\varnothing)$，用于估计高斯后验的概率分布。

3.4 生成模型

生成网络，由参数 θ 参数化，对潜在动态（$\mathbf{h}_{1:T}$，$\mathbf{s}_{1:T}$）以及生成过程（$\mathbf{o}_{1:T}$，$\mathbf{y}_{1:T}$，$\mathbf{a}_{1:T}$）进行建模。它包括一个门控循环单元 $\mathcal{F}_\theta$，一个先验网络（$\mu_\theta, \sigma_\theta$），一个图像解码器 $\mathcal{G}_\theta$，一个鸟瞰图解码器 $\mathcal{L}_\theta$，和一个策略 π_θ。

先验网络估计高斯分布的参数 $p(s_t|h_{t-1}, s_{t-1}) \sim \mathcal{N}(\mu_\theta(h_t, \hat{a}_{t-1}), \sigma_\theta(h_t, \hat{a}_{t-1})I)$，其中 $h_t = \mathcal{F}_\theta(h_{t-1}, s_{t-1})$，$\hat{a}_{t-1} = \pi_\theta(h_{t-1}, s_{t-1})$。由于先验不具有对 a_{t-1} 的真实行动信息，后者用学到的策略 $\hat{a}_{t-1} = \pi_\theta(h_{t-1}, s_{t-1})$ 进行估计。

3.5 模仿学习

该论文使用行为克隆来模仿人类专家数据

3.6 强化学习

我们引入了具有特权模型的强化学习教练，它可以利用多种传感器信息来感知环境并输出结果。其强大的监督能力提升了模仿学习在某些场景下的性能。

4 复现细节

4.1 与已有开源代码对比

该论文使用模仿学习来模仿人类专家数据可能会使智能体在某些场景下表现不好。为此，我引入了具有特权模式的强化学习教练来提升其性能，它可以利用多种传感器信息来感知环境并输出结果。实验结果表明其在一定程度上有性能提升。图 2 为我们的改进示意图

图 2. 改进示意图

4.2 实验环境搭建

本实验基于 ubuntu20.04 进行实验，使用 8 张 3090 作为 gpu。在 carla 仿真上执行 leaderboard 任务。

4.3 创新点

该论文使用模仿学习来模仿人类专家数据可能会使智能体在某些场景下表现不好。为此，我引入了具有特权模式的强化学习教练来提升其性能，它可以利用多种传感器信息来感知环境并输出结果。实验结果表明其在一定程度上有性能提升。

5 实验结果分析

从实验数据可以看出具有特权模式的强化学习教练可以很好地提升其性能。在 carla 的 leaderboard 任务上奖励有所提升，碰撞率有所下降。

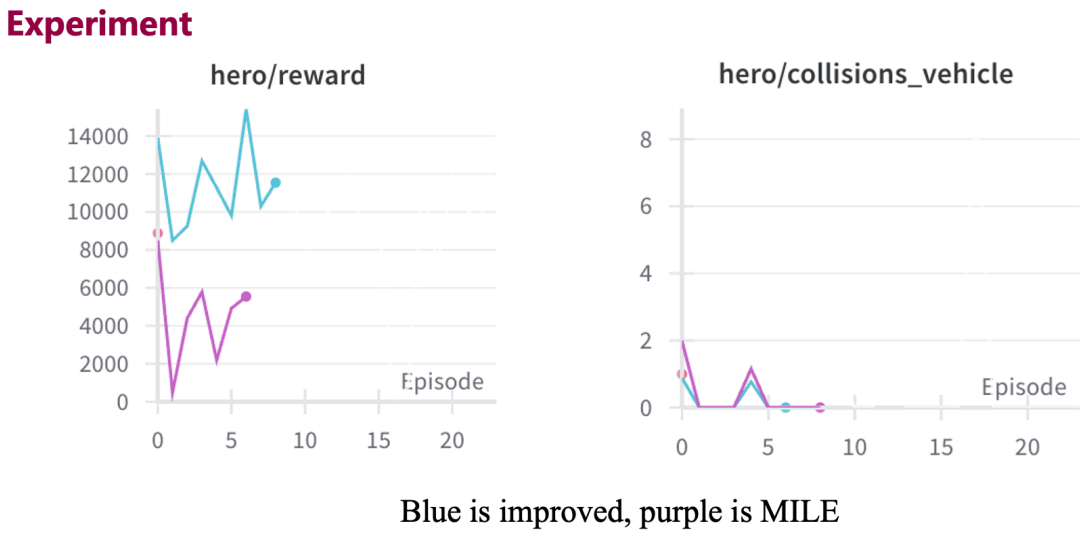

图 3. 实验结果

6 总结与展望

该论文提出了一种基于模型的模仿学习方法，用于共同学习世界模型和自主驾驶的策略。该论文利用三维几何作为归纳偏置，直接从专家演示的高分辨率视频中学习高度紧凑的潜在空间。该模型可以预测各种合理的状态和行动，这些状态和行动可以被解释为鸟瞰式的语义分割。此外，它可以根据完全在想象中预测的计划执行复杂的驾驶动作。但是该论文使用模仿学习来模仿人类专家数据可能会使智能体在某些场景下表现不好。为此，我引入了具有特权模式的强化学习教练来提升其性能，它可以利用多种传感器信息来感知环境并输出结果。实验结果表明其在一定程度上有性能提升。

在未来，我将沿着这个方向，需要专注于考虑如何在特殊场景下提升智能体性能。

参考文献

[1] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017.

[2] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[3] Mariusz Bojarski, Philip Yeres, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Lawrence Jackel, and Urs Muller. Explaining how a deep neural network trained with end-to-end learning steers a car. *arXiv preprint arXiv:1704.07911*, 2017.

[4] Silvia Chiappa, Sébastien Racaniere, Daan Wierstra, and Shakir Mohamed. Recurrent environment simulators. *arXiv preprint arXiv:1704.02254*, 2017.

[5] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International conference on machine learning*, pages 1174–1183. PMLR, 2018.

[6] Peter Englert, Alexandros Paraschos, Marc Peter Deisenroth, and Jan Peters. Probabilistic model-based imitation learning. *Adaptive Behavior*, 21(5):388–403, 2013.

[7] Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction. In *International Conference on Machine Learning*, pages 3233–3246. PMLR, 2020.

[8] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

[9] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019.

[10] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

[11] Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen, Vinh-Dieu Lam, Alex Bewley, and Amar Shah. Learning to drive in a day. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8248–8254. IEEE, 2019.

[12] Rahul Kidambi, Jonathan Chang, and Wen Sun. Mobile: Model-based imitation learning from observation alone. *Advances in Neural Information Processing Systems*, 34:28598–28611, 2021.

[13] Markus Kuderer, Shilpa Gulati, and Wolfram Burgard. Learning driving styles for autonomous vehicles from demonstration. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 2641–2646. IEEE, 2015.

[14] Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search of static linear policies is competitive for reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[15] Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988.

[16] Dean A Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural computation*, 3(1):88–97, 1991.

[17] Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279, 2019.

[18] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[19] Sahand Sharifzadeh, Ioannis Chiotellis, Rudolph Triebel, and Daniel Cremers. Learning to drive using inverse reinforcement learning and deep q-networks. *arXiv preprint arXiv:1612.03653*, 2016.

[20] Niklas Wahlström, Thomas B Schön, and Marc Peter Deisenroth. Learning deep dynamical models from image pixels. *IFAC-PapersOnLine*, 48(28):1059–1064, 2015.

[21] Manuel Watter, Jost Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. *Advances in neural information processing systems*, 28, 2015.

[22] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

[23] Wenxuan Zhou, Sujay Bajracharya, and David Held. Plas: Latent action space for offline reinforcement learning. In *Conference on Robot Learning*, pages 1719–1735. PMLR, 2021.