



2018年深度强化学习研讨会

基于深度强化学习 的礼让自动驾驶研究

Deep Reinforcement Learning and Reciprocal Automatic Driving

杨明珠
大连交通大学

吴焦苏 李真真
中国科学院人工智能联盟标准组
中国科学科技战略研究院

2018.8.4

目录

深度强化学习理论

自动驾驶技术现状与问题

深度强化学习在自动驾驶技术中的应用

基于深度强化学习的礼让自动驾驶研究

结论与展望

深度强化学习理论

DQN

算法: Deep Q Networks

模型:

$$Q^*(s, a) = \mathbb{E}_{s' \sim \mathcal{E}} \left[r + \gamma \max_{a'} Q^*(s', a') \mid s, a \right]$$

$$L_i(\theta_i) = \mathbb{E}_{s, a \sim \rho(\cdot)} \left[(y_i - Q(s, a; \theta_i))^2 \right]$$

$$\nabla_{\theta_i} L_i(\theta_i) = \mathbb{E}_{s, a \sim \rho(\cdot); s' \sim \mathcal{E}} \left[\left(r + \gamma \max_{a'} Q(s', a'; \theta_{i-1}) - Q(s, a; \theta_i) \right) \nabla_{\theta_i} Q(s, a; \theta_i) \right]$$

适用于离散情况，并且变体较多

V Mnih, K Kavukcuoglu, et al. Playing Atari with Deep Reinforcement Learning, Computer Science, 2013.

A3C

算法: Asynchronous Advantage Actor-Critic

模型:

$$L_i(\theta_i) = \mathbb{E} \left(r + \gamma \max_{a'} Q(s', a'; \theta_{i-1}) - Q(s, a; \theta_i) \right)^2$$

$$g = \alpha g + (1 - \alpha) \Delta \theta^2 \text{ and } \theta \leftarrow \theta - \eta \frac{\Delta \theta}{\sqrt{g + \epsilon}},$$

Asynchronous one-step Q-learning

Asynchronous one-step Sarsa

Asynchronous n-step Q-learning

Asynchronous advantage actor-critic

V Mnih, AP Badia, et al. Asynchronous Methods for Deep Reinforcement Learning, arXiv:1602.01783.

DDPG

算法： Deep Deterministic Policy Gradient

• 模型：

$$\begin{aligned}\nabla_{\theta^\mu} J &\approx \mathbb{E}_{s_t \sim \rho^\beta} \left[\nabla_{\theta^\mu} Q(s, a | \theta^Q) \Big|_{s=s_t, a=\mu(s_t | \theta^\mu)} \right] \\ &= \mathbb{E}_{s_t \sim \rho^\beta} \left[\nabla_a Q(s, a | \theta^Q) \Big|_{s=s_t, a=\mu(s_t)} \nabla_{\theta^\mu} \mu(s | \theta^\mu) \Big|_{s=s_t} \right]\end{aligned}$$

$$\mu'(s_t) = \mu(s_t | \theta_t^\mu) + \mathcal{N}$$

TP Lillicrap, JJ Hunt, et al. Continuous control with deep reinforcement learning, Computer Science , 2015 , 8 (6) :A187.

D Silver, G Lever, et al. Deterministic policy gradient algorithms, International Conference on International Confe..., 2014 :387-395.

TRPO

算法: Trust Region Policy Optimization

模型:

$$\underset{\theta}{\text{minimize}} [L_{\theta_{\text{old}}}(\theta) + CD_{\text{KL}}^{\text{max}}(\theta_{\text{old}}, \theta)]$$

$$\underset{\theta}{\text{minimize}} L_{\theta_{\text{old}}}(\theta)$$

$$\text{subject to } D_{\text{KL}}^{\text{max}}(\theta_{\text{old}}, \theta) \leq \delta.$$

$$\underset{\theta}{\text{minimize}} L_{\theta_{\text{old}}}(\theta)$$

$$\text{subject to } \overline{D}_{\text{KL}}^{\rho_{\theta_{\text{old}}}}(\theta_{\text{old}}, \theta) \leq \delta.$$

TRPO

技巧一：对状态分布进行处理

技巧二：利用重要性采样对动作分布进行的处理

技巧三：在约束条件中，利用平均KL散度代替最大KL散度

J Schulman, S Levine, et al. Trust Region Policy Optimization, Computer Science ,
2015 :1889-1897.

PPO

算法: Proximal Policy Optimization

模型:

$$L_t^{CLIP+VF+S}(\theta) = \hat{\mathbb{E}}_t[L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\theta) + c_2 S[\pi_\theta](s_t)]$$

$$\hat{A}_t = -V(s_t) + r_t + \gamma r_{t+1} + \dots + \gamma^{T-t+1} r_{T-1} + \gamma^{T-t} V(s_T)$$

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \dots + (\gamma\lambda)^{T-t+1}\delta_{T-1}$$

$$\text{where } \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$$

OpenAI 目前默认深度强化学习算法

PPO 在样本复杂性和易于调优之间取得平衡, 试图在每一步最小化成本函数计算更新时, 确保与先前策略的偏差相对较小。

J Schulman, F Wolski, et al. Proximal Policy Optimization Algorithms,
arXiv:1707.06347.

HER

算法：Hindsight Experience Replay，也叫“事后诸葛亮”

模型：

$$Q^*(s, a) = \mathbb{E}_{s' \sim p(\cdot | s, a)} \left[r(s, a) + \gamma \max_{a' \in \mathcal{A}} Q^*(s', a') \right]$$

在完成了一次动作之后再选定目标、重放经验进行学习，即从错误中进行学习。

M Andrychowicz, F Wolski, et al. Hindsight Experience Replay,
arXiv:1707.01495.

自动驾驶技术现状与问题

自动驾驶技术流程图



自动驾驶公司

互联网公司:

Google(Waymo), 百度, 苹果, Uber等。



自动驾驶公司

传统车企：

福特， 博世， 大众， 通用， 宝马， 奔驰等。



自动驾驶技术的三大问题

感知问题：

信息预处理过程是感知问题的重要体现，传感器检测到物体信息，并进行检测，分割与识别，激光雷达测量物体距离，摄像头和相机记录视觉信息，并将所有的信息进行整合与处理。

对图像信息采取语义分割，使用VGG算法对物体进行识别，使用YOLO算法对摄像头的信息进行检测。将处理后的信息传送给决策单元。但是一天中不同的时间段、环境背景和任何可能的运动都会对感知部分做出影响；而且传感器所采集的各数据类型之间的差异，确认物体的存在性及其类型所需的传感器融合算法在技术上实现起来是极具挑战性的。

在极端天气或者光线问题不同的情况下，以及障碍物遮挡等问题出现时，视觉检测技术如何达到良好的检测与识别效果。

自动驾驶技术的三大问题

- 决策问题：
- 为了模仿人类的决策，自动驾驶车辆必须历经大量的应用情景并在不同的场地内进行密集且全面的“训练”。
- 深度强化学习就可以很好的解决决策方法的问题，首先，如何通过Actor-critic机制选择最优policy；

其次，如何将经验全部存储在Replay Buffer当中，有利于直接提取训练参数。

最后，如何采用异步的方式进行多目标进行训练，节约时间和成本。

自动驾驶技术的三大问题

控制问题：

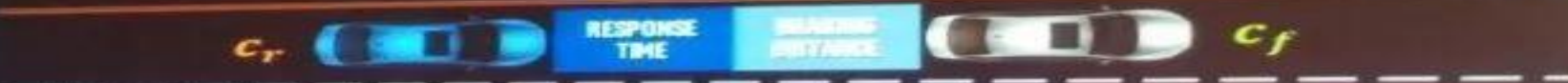
系统还需要一个故障安全机制（**Fail-Safe Mechanism**），该机制能确保在汽车发生故障时不会让车上的乘客和周围的人员陷于险地。

目前尚无方法来检查每一个可能的软件状态及其所造成的结果，建立防护措施以防止最坏结果的发生，同时控制车辆以使其安全地停车仍是待解决的难题。因此，冗余设计和长时间的测试工作将是必须的。

自动驾驶技术的三大问题

Mobileye的RSS模型

DEFINE SAFE LONGITUDINAL DISTANCE

$$d_{min} = \left[v_r \rho + \frac{1}{2} \alpha_{max} \rho^2 + \frac{(v_r + \rho \alpha_{max})^2}{2\beta_{min}} - \frac{v_f^2}{2\beta_{max}} \right]_+$$


α_{max} Max acceleration during response time (for c_r)

β_{max} Max braking applied by c_f

不主动：事故的发生不是自动驾驶汽车引起的；

不拒绝：你非要把我拉入到事故进行亲密接触，我也没辙；

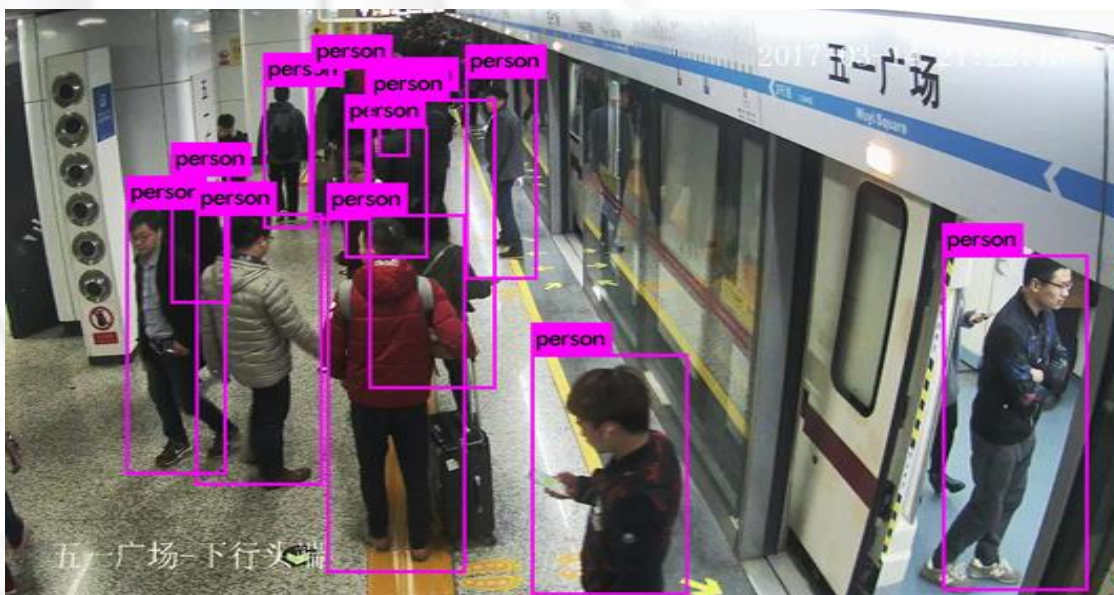
不负责：有了不主动，不拒绝，自然也就不负责。

--王宏明《再见吧春天|自动驾驶的冬天》

解决自动驾驶技术问题的两种方法

方法一：

低精度定位+低精度地图+高准确识别率

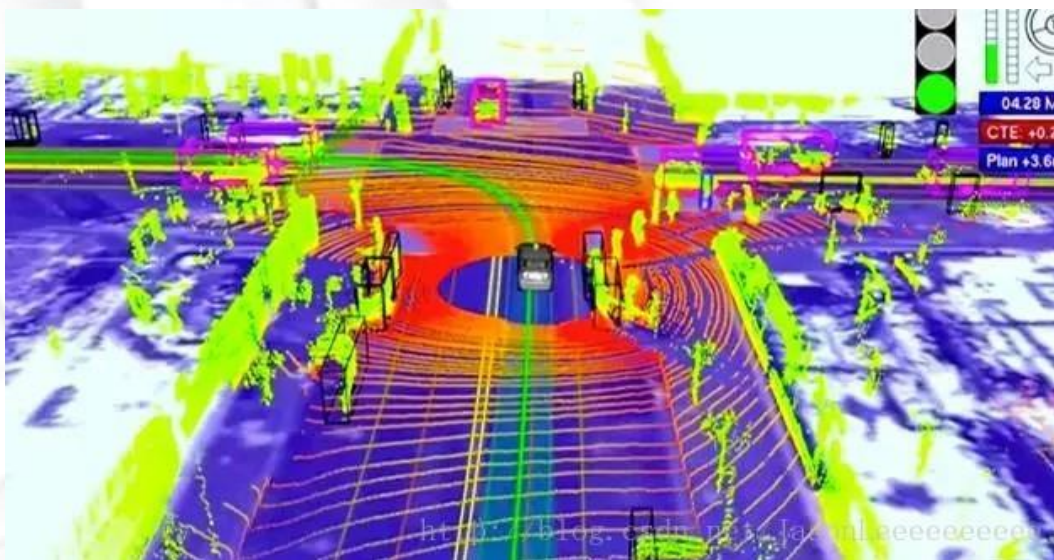


基于YOLO算法的地铁人员检测与识别
深度强化学习目前主要应用于方法一

解决自动驾驶技术问题的两种方法

方法二：

高精度定位+高精度地图+更准确的识别率



激光点云制作高精度地图

深度强化学习未来可以应用在方法二中，提高识别率

DeepMap

DeepMap提供软件，这些软件可以帮助汽车绘制高清图像，完成实时定位，它还提供必要的基础设施，让技术可以大规模应用于汽车。DeepMap相当于无人驾驶汽车的眼睛和大脑，它给环境绘图，告诉汽车它是否与环境存在关联，然后将信息通过DeepMap平台与其它汽车分享。

DeepMap的目标是帮助无人驾驶汽车和无人驾驶卡车安全行驶，让无人驾驶尽快变成现实。要达到目标，需要精准的绘图和定位技术，必须重新开始打造，确保安全、性能、运行效率达到最高等级。要在真实世界安全部署相当困难，很费时间，成本也很高，这些挑战需要解决。

深度强化学习在自动驾驶技术中的应用

采用深度强化学习的自动驾驶团队

WAYVE团队（英国剑桥大学两位机器学习博士创立的英国自动驾驶汽车公司）

本田研究院团队（宾夕法尼亚大学、本田研究院和乔治亚理工学院合作）

堪萨斯州立大学团队（Kansas State University, KSU）

韩国汉阳大学团队（机器监测和控制实验室博士生）

Wayve团队

问题：如何使自动驾驶车辆从头开始学习保持在同一车道内行驶。

算法：DDPG

成果：使用连续性动作算法DDPG，训练自动驾驶车辆沿同一车道内行驶，团队仅利用单路摄像头来让自动驾驶车辆学习，并且在30分钟内，自动驾驶车辆学会保持在同一车道内行驶250米。

缺点：没有感知的预处理过程和控制单元的预警与防碰撞系统。

Alex K, Jeffery H, et al. Learning to drive in a day, arXiv:1807.00412.

本田研究院团队

问题：如何确定驾驶员意图，提供安全导航的有效策略通过无信号的交叉路口。

算法：DQN

成果：在通过交叉路口时，与 TTC (Time-to-Collision|碰撞时间)相比，DQN 方法在实现目标上要有效得多。DQN 可以准确预测远处车道在当前车辆通过该车道时的交通状况；该 DQN 司机还能预测即将到来的车流是否有足够的时间制动。

缺点：仅适用于离散型动作的决策任务，不适合在连续性动作中使用。无控制单元的预警和防碰撞系统。

D. Isele , et, al. Navigating Occluded Intersections with Autonomous Vehicles using Deep Reinforcement Learning, arXiv:1705.01196.

堪萨斯州立大学团队

问题：如何建立一个基于深度强化学习的新框架在对抗性极强的情况下，对碰撞避免机制的行为进行训练，使系统进入不安全预警状态.

算法：DDPG,TRPO,A3C

成果：提出了一种基于深度强化学习的新框架，用于对自动驾驶汽车的碰撞避免机制的行为进行基准测试。

缺点：无感知单元的预处理过程，并且没有在连续性动作的决策任务。

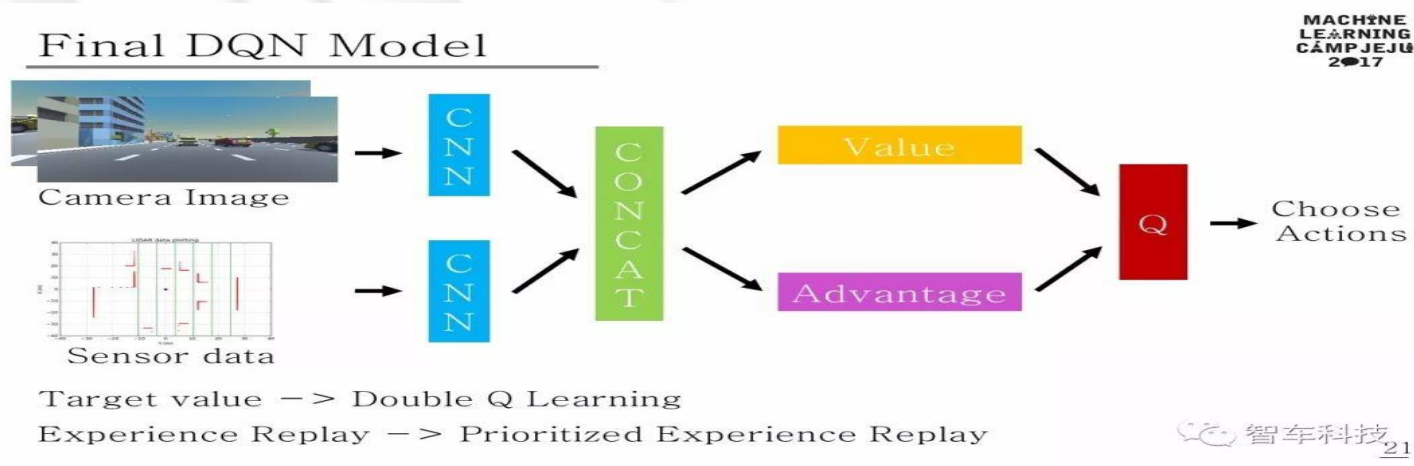
Vahid B, Arslan M, et al. Adversarial Reinforcement Learning Framework for Benchmarking Collision Avoidance Mechanisms in Autonomous Vehicles, arXiv:1806.01368.

韩国汉阳大学团队

问题：在车辆密集的交通行驶时奖励机制的设计：1.找到使车辆高速行驶的策略，2.以无碰撞的轨迹行驶，3.不频繁地改变车道。

算法：DQN

模型：



- Kyushik Min. Deep Q Learning Based High Level Driving Policy Determination, Hayoung Kim and Kunsoo Huh, Member, IEEE.

韩国汉阳大学团队

成果：提出了一个使用传统DAS和深度强化学习融合的自动驾驶框架。该框架在DAS(驾驶辅助系统)功能（例如车道变换，巡航控制和车道保持等）下，使用深度强化学习算法来训练车辆，在模拟高速公路场景中成功驾驶，使用多模式输入的驾驶策略网络，不会造成不必要的车道变化，最大限度地提高平均速度，和最少车道变化为规则，来确定超车次数，车辆比具有单输入的车辆更好地驾驶，大多数情况下可以保证车辆行驶的安全。自主车辆可以由受过深度强化学习训练的主管来控制。

缺点：仅适用于离散型动作的决策任务，不适用于连续性决策任务。

基于深度强化学习的礼让自动驾驶研究

本团队综合了国际上最先进的多个自动驾驶团队的优缺点，整合其优秀思想与先进算法，从感知单元，决策单元，控制单元统一研究基于深度强化学习的礼让自动驾驶技术。

“安全行车，礼让三先”，礼让三先：先让、先慢、先停。绝对不可抢行争路，互不相让，以致形成僵持局面。根据我国《道路交通安全法》及相关规定及《道路通行规定》，一般情况下的会车，须遵守下列规则：空车让重车，单车让拖挂货车，大车让小车，货车让客车，教练车让其他车辆，普通车让执行任务的特种车，下坡车让上坡车；机动车让非机动车，非机动车让行人。

我们根据这一思想，提出礼让自动驾驶概念，自动驾驶车辆主动进行避让，无论是车辆自身的的撞击过程还是别人对我们的撞击都主动采取避让措施。

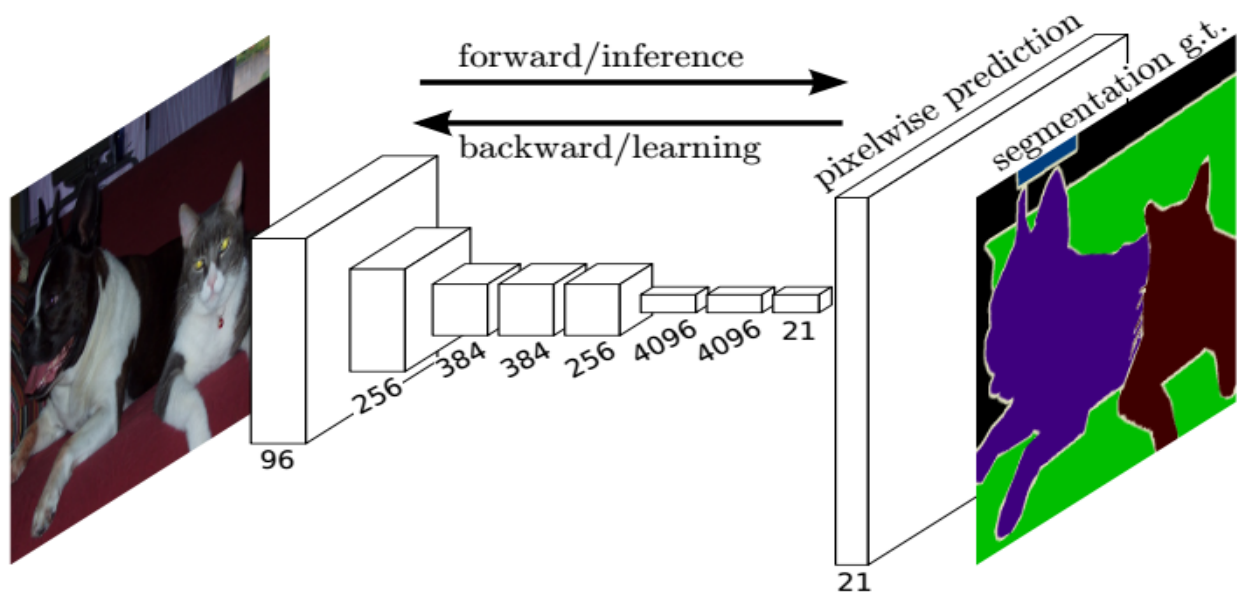
感知

在车辆感知部分加入交通标识识别，通过深度卷积神经网络(DCNN)中的语义分割算法(FCN)，将道路无关信息分离出去，只留下对车辆行驶有用的部分指导车辆行驶；通过对双路摄像头采集到的图像信息进行分割，识别和检测等预处理过程将信息传递给决策单元进行之后的决策与控制功能。

在极端天气或者光线问题不同的情况下，以及障碍物遮挡等问题出现时，目前深度卷积神经网络VGG可以对车辆识别正确率达到96%以上，并且在行人识别问题上也表现出良好的效果，但是部分极端天气下依然存在降低检测结果的问题存在。但是目前的机器视觉技术受限于硬件和其他极端情况的影响，短期内识别效果不会有更好的提升。

感知

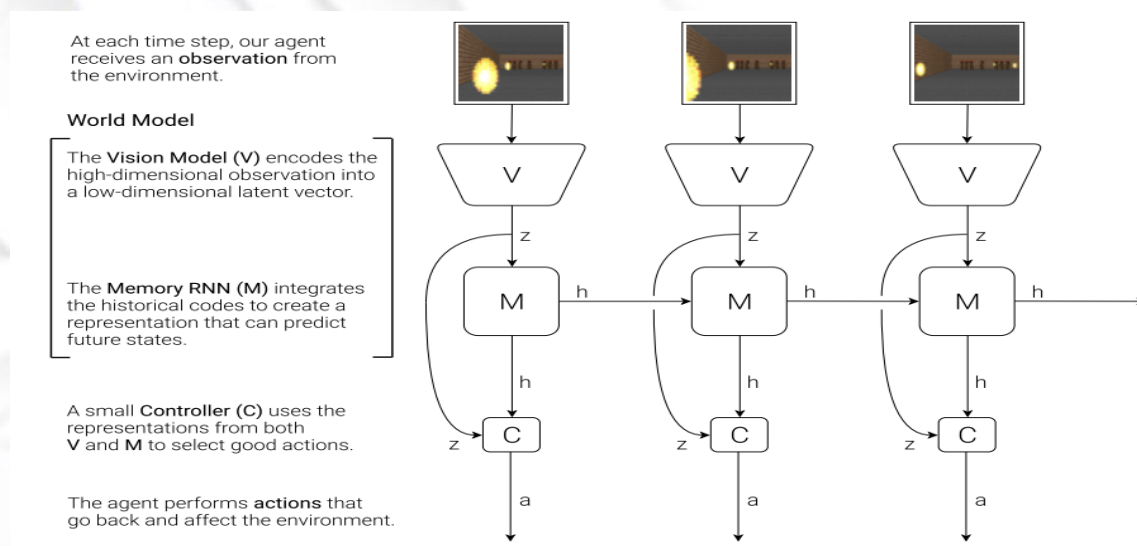
分割采用实例分割（instance segmentation）和语义分割（semantic segmentation）等方法，或者使用全景分割（panoptic segmentation）方法来进行图像分割；检测技术我们主要采用YOLO方法；识别方面主要采用VGG的模型来做图像识别方面的工作。



感知

感知模块将地图、三维传感器、二维传感器中的信息给到“世界模型”（**world model**），世界模型将上述信息，汇总在一张地图中，理解每一个时刻不同的物体相对于路面、道线等的位置，预测下一刻的可选路径都有哪些。

世界模型整体架构：



D Ha , J Schmidhuber. World models. arXiv:1803.10122.

决策

使用DDPG算法思路改进自动驾驶决策部分，加入礼让自动驾驶的部分，即在动态连续性动作的情况下使得自动驾驶车辆避免发生碰撞。

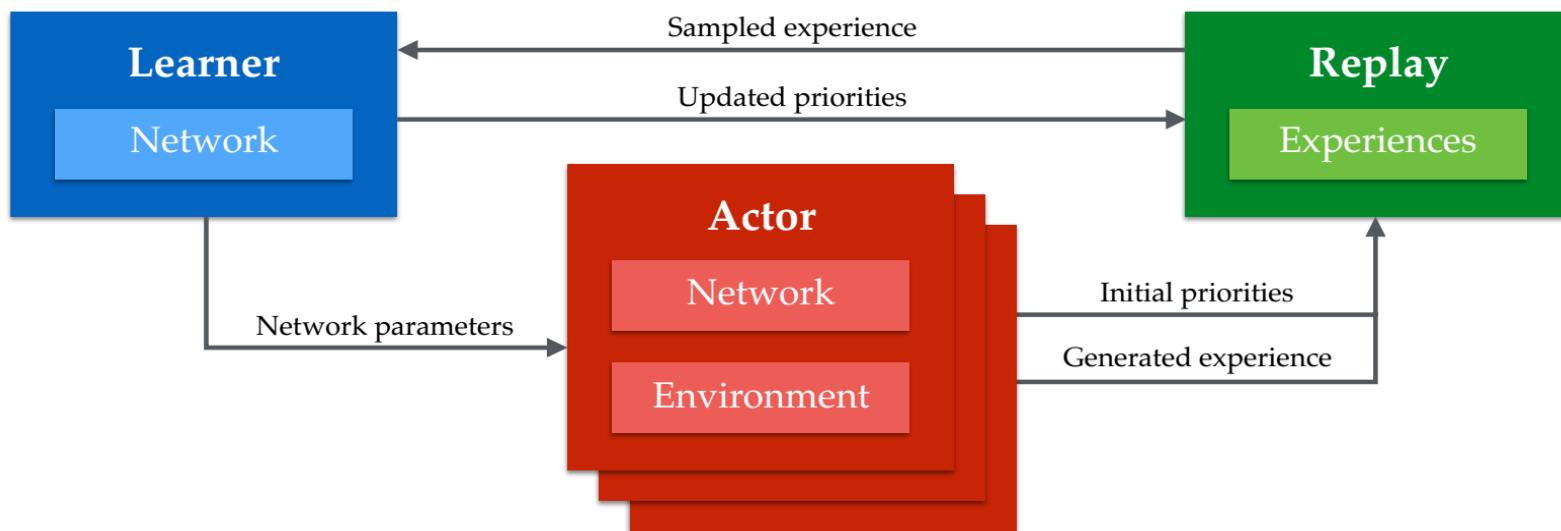
$$P(N_i|x) = \binom{n}{N_{i,1}, N_{i,2}, \dots, N_{i,k}} \prod_{j=1}^k x_j^{N_{i,j}}$$
$$f_i(x) = \frac{\sum_j P(N_j|x) U_{j,i}}{1 - (1 - x_i)^k}$$

难点：“礼让”的概念在机器人领域中应用较多，车辆在高速行驶过程中的礼让问题就会有很大的技术难度。

决策

结合PPO与HER算法的思想，PPO提升算法的效率，提高Policy选择的准确率，HER促使自动驾驶车辆在每次训练之后总结经验，在下一次训练中不犯同样的错误。

使用深度强化学习不但可以做决策会比一般的深度学习算法(如RNN和LSTM)拥有更好的决策能力，Actor-Critic机制的使用对policy的选择几乎达到最优,Replay Buffer存储学习过的所有policy与经验值，在之后有相似动作与行为决策过程中直接提取policy，不需要再去进行训练与学习。



控制

通过图像分割的方法分离车道线，当自动驾驶车辆偏离车道线时的距离检测，同时分割出前方车辆，给出车距监测结果，避免碰撞时间(TTC)，防前碰撞方向和防后碰撞方向，使得自动驾驶车辆可以在道路上安全的行驶。

采用双保险安全机制，当决策发生错误时，或有突发情况时，车辆自动制动。

激光雷达检测周围出现的车辆，对车辆的距离进行测量，若车辆进入安全行车区域内，自动驾驶车辆减速或制动。

仿真

语言：Python

框架：Tensorflow，Keras

仿真平台：TORCS(The Open Racing Car Simulator)

简介：一款开源3D赛车模拟游戏，在Linux操作系统上广受欢迎。该平台拥有50种车辆和20条赛道，视觉效果简单明了，但该仿真环境较简单，不适合做复杂场景下的自动驾驶仿真。

结论与展望

结论与展望

自动驾驶技术应用的几种深度强化学习算法总结：

DQN出现最早，改良版本最多，离散情况效果最佳，原理相对较简单，易于掌握与入门。

DDPG是在**DQN**的基础上进行改良，原理易懂，在连续动作中表现优异，适用于自动驾驶系统的决策研究。

之后出现的**A3C**,**PPO**,**HER**等算法在连续动作中都有很好的应用与体现。

目前，有很多人在将分层强化学习和逆向强化学习（模仿学习）应用于自动驾驶技术当中，效果有待考究实验。

结论与展望

实际上，基于时间空间的博弈动力学研究表明，机器人在目前的实验与发展状态下不具备伦理判断能力与决策功能。所以，将机器人置于伦理困境是超出了机器人研究的能力范围。

德国联邦交通与数字基础设施部伦理委员会确认：自动驾驶系统需要更好的来适应人之间的交流习惯，并不是需要人来更好的适应机器，这也是我们开发机器人的原因；开发者与立法者必须清晰界定机器和人之间的责任，同时，鉴定机器不能取代或优先于人的自主决定权的立场，面对不可避免的事故，最终的行为决定权必须要由人掌握。

--吴焦苏 《未来法治研究院网络法第十二期读书会》

结论与展望

自动驾驶系统的设计只有在立法者确定在决定权从机器向人的顺利转移的情况之下，自动驾驶车辆才可以被批准上路。

人类对这些问题所产生的技术上的困难应当有清醒的认识，在自动驾驶车辆正式上路之前，开发者及生产商应当进行充分的破坏性试验和严格的封闭路测，在保证其在各种情况下都可以有更好的表现的前提下才可以进行路测等试验。在自动驾驶系统的安全性不能得到严格的保证之前，自动驾驶系统不应当被批准量产。

--吴焦苏 《未来法治研究院网络法第十二期读书会》

欢迎批评指正!