

深度强化学习在自动驾驶技术中的应用——杨明珠

读芯术 大数据创新学习中心 8月16日

不到现场，照样看最干货的学术报告！

这里是学术报告专栏，与读芯术同步报道“2018深度强化学习：理论与应用”研讨会的系列报告。

2018年8月4日，北京理工大学大数据创新学习中心与中国科学院人工智能联盟标准组联合主办了为期一天的专家讲座活动-----“2018深度强化学习：理论与应用”学术研讨会。活动现场参与人数超过600人，在线同步观看人数超过12万人。学界与业界专家齐聚一堂，共同分享学习深度强化学习领域的最新研究成果。本文小编亲临现场，为您揭秘深度强化学习在自动驾驶技术中的应用详细报告。活动视频在文末，相关报告ppt也在文中同步分享。

深度强化学习在自动驾驶技术中的应用



杨明珠 大连交通大学

今天我的演讲内容主要分为四个部分：深度强化学习的理论、自动驾驶技术的现状以及问题、深度强化学习在自动驾驶技术当中的应用及基于深度强化学习的礼让自动驾驶研究。

首先是深度强化学习的理论，DQN做了深度的拓展，在离散型动作中应用效果比较好，但连续性动作当中表现效果并不好，所以做了一些改进和发展，如Double DQN等。

在连续型动作之中我个人比较喜欢DDPG的理念，原因有两点：①之前学习到的经验和Policy数据放到Replaybuffer当中，若之后的行为当中发现和之前相似的地方就会直接从Replaybuffer当中把之前的经验和数据直接调用出来，这样就可以避免在重复进行一种训练或者采集的方式，节省时间、提高效率；②信任域的策略优化，简称TRPO，其实是对之前的算法做了改进，如对状态分布进行处理，利用重要性采样对动作分布进行的处理及在约束条件当中，把平均KL散度代替最大KL散度。

PPO也是最近比较热门的一种深度强化学习算法，分为N个Actor，同时进行一些工作，这样平均分配给很多个actor，合作来做的话效率会更高，而且会节省更多的时间。HER算法也是个人最喜欢的之前经过所有训练，经验总结出来，这个工作结束以后全部消化一遍，然后做第二次实验或者工作的时候吸取了前面的经验，然后再进行下面的训练或者工作的话，就会避免一些错误，如无人驾驶撞车了，上次为什么撞车了呢？第二次需要避免这个错误，即不让它撞车。

自动驾驶技术的现状和问题，主要成为三个模块：①感知模块，包括摄像头、传感器，即硬件方面，采集到的图像信息、视频信息或者传感器的数据反馈到了决策模块，也叫黑匣子，是自动驾驶技术当中具有决定性的模块，主要包括Planning和之后的预测。②决策模块，主要包括GPU、CPU等计算单元。③控制模块，主要是对自动驾驶的控制，比如制动和减速。预警系统，如果是突发情况，决策模块反应不过来就会直接给到预警系统，采取制动或者减速。

自动驾驶公司分为互联网公司（如Google、百度、苹果和Uber）及传统车企（如福特和汽车配件的博世、大众、通用、宝马和奔驰等）。**目前自动驾驶技术有三个问题：**①感知方面也可以叫做信息的预处理，主要包括对图像或者视频信息的分割、检测或者识别，如果识别的准确率更高可能会对之后的决策有比较好的优势。运行当中也需要用到分割工作，如沿着车线走需要分割车线位置等。②决策方面其实是为了模仿人类，所以需要经过很多训练，利用强化学习来做自动驾驶即像人考驾照的过程，学习怎样开车，最后达到上路的水平。③控制方面就是故障安全机制，遇到危险的情况下来不及反应，就需要安全机制保障车内的人身安全，我们做自动驾驶也就是为了减少交通事故的发生率，让更多的人可以安安全全地坐上自动驾驶汽车。

在控制方面，我们不得不提一个模型，就是Mobileye的RSS模型。我们在此借用了王宏明教授的一段话：不主动、不拒绝、不负责，事故发生不是自动驾驶汽车引起的，非要拉进事故亲密接触也没有办法，有了不主动、不拒绝，自然也就不负责了。RSS模型就是不去主动撞你，但你要是撞我的话我没有任何办法。我们不能说它不好，但确实是给了我们一种启示，就是不要去撞别人，别人来撞我们就真的置之不理吗？王宏明教授批评也是因为这个原因，别人撞我的时候我没有任何反

应，这样的想法也是不对的。RSS模型也有可以肯定的地方，确实是做到了不去撞别人，之后的工作也是在RSS模型上面做一些改进。

现在解决自动驾驶技术问题有两种方法：一种是低精度定位+低精度地图+高准确识别率，另一种是高精度定位+高精度地图+更准确的识别率。

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



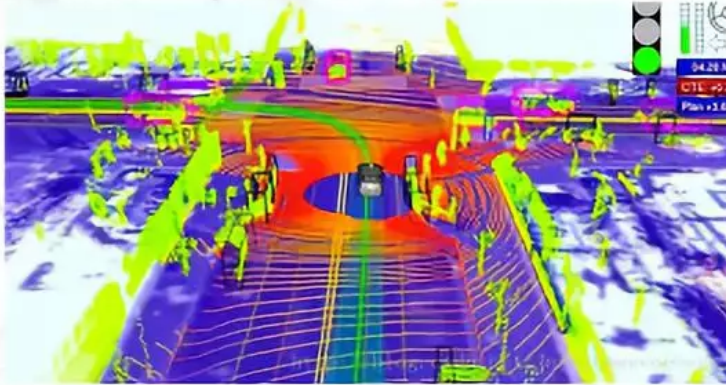
基于YOLO算法的地铁人员检测与识别

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

26

国内百度采用的是高精度的定位和高精度的地图，能够知道是在哪条路上面，包括之后的Planning也是依据高精度地图来做规划。个人觉得这种方式可能比较麻烦，因为对要求会特别高，如果高精度地图有一点偏差的话可能对之后的决策和规划会有一些麻烦，所以我们将采用深度强化学习算法来根据高识别率的信息做决策，不太依赖于高精度地图。

DeepMap公司也在做高精度地图方面的研发工作，目标就是用无人驾驶汽车有眼睛、有大脑，而且可以确保更加安全地到达之后想要去的地方，但这样也是非常费时间、成本非常高，因为需要各条街路采集信息，包括全景。

百度是有采集信息的车辆，其实也是比较辛苦的，需要采集所有全景的图像来做上传，最后再和百度地图结合，这样才能制定比较好的高精度地图，这样成本会非常的高。

关于深度强化学习在自动驾驶当中的应用，有几个团队：WAYVE团队、本田研究院团队、堪萨斯州立大学团队、韩国汉阳大学团队。Wayve是我个人比较欣赏的团队，是由英国剑桥的博士毕业生创立的自动驾驶。

Wayve在今年7月发布的文章是《**Learning to Drive in a Day**》，仅仅用了一个前景摄像头，就是车前方的视频作为输入的State，输出的Action就是保证在同一车道内行进距离，行驶距离长，reward就大；行驶距离短，reward就短。结果是只用了单个摄像头让自动驾驶汽车在三十分分钟内

学会了保持在同一车道内行驶二百五十米距离。这样的方式我们是比较欣赏，但不太建议使用这种仅仅基于视觉的方式来做自动驾驶，因为开车肯定是眼观六路耳听八方，侧面或者后面出现任何问题没有办法及时预警，没有办法及时处理，将来在上路的问题上肯定是有很大的缺陷。

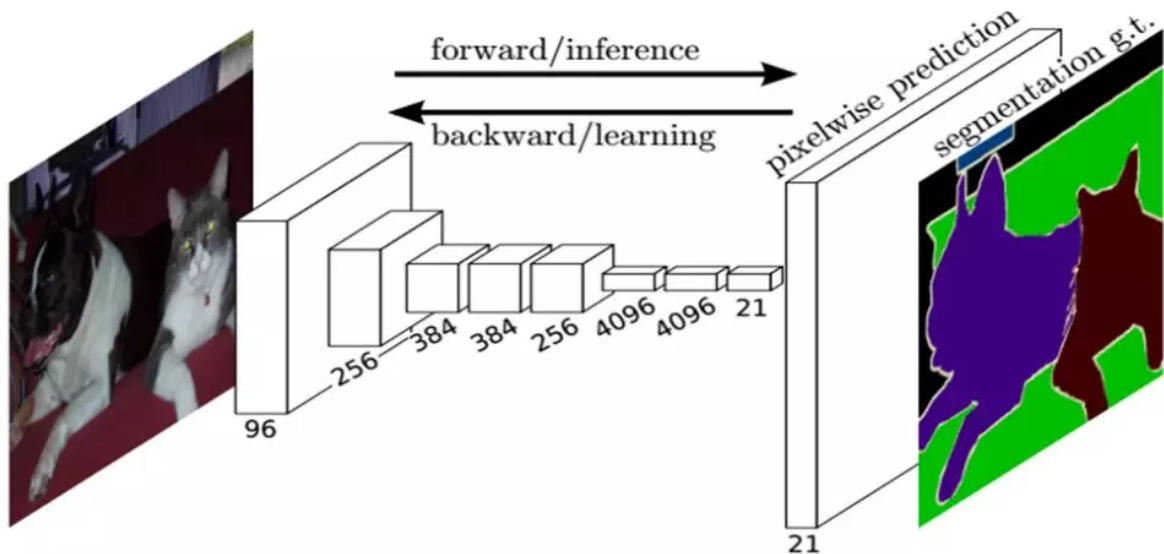
宾夕法尼亚大学，本田研究院和乔治亚理工学院合作团队是采用TTC模式，能够提前知道岔路口的状态，如何通过岔路口并且预测到达这个岔路口的时间是不是有危险，有没有足够的时间进行制动，TTC一般都是二点七秒，那个，该团队存在的缺陷因为就是DQN存在的问题，即在离散动作当中表现优异，在连续性动作中表现不好，如果是在高速行驶的情况下如何应用，解决得并不是太好。

如何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，堪萨斯州立大学团队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，但是有一个缺点：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

韩国汉阳大学团队使用了传统的驾驶辅助系统和DQN结合，是在高速公路的模拟场景当中驾驶，而且采用的是两个DQN的模式作为输入，看一看是不是在车道变化情况下有一个很好的效果，超车的时候是不是也会有更好的效果，存在的问题其实也就是DQN的问题，离散性会更好一些，连续性并不是太好。

我们将这些思想做了融合，提出了我们的一种新的自动驾驶技术，就是礼让自动驾驶。**我们的礼让自动驾驶也是从三个方面来说：感知、决策和控制单元。**什么叫做礼让，包括“安全行车、礼让三先”：先让，先慢，先停，我不去撞别人，别人撞我的时候要先避让一下，避免发生撞击的情况。

感知部分是围绕检测、识别和图像分割等方面，检测当中我们用的最多的是YOLO算法，如果车速特别快的话也需要快速的检测，然后再去做一些决策方面的，识别方面个人比较喜欢VCG模型，模型结构简单而且，识别效果也是比较不错的。分割当中有局部分割、语义分割和全景分割，现在应用最多的是语义分割和全景分割。感知模块我们借鉴AndreasGeiger的思想，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），我们把感知部分所有信息汇总到一个地图当中，做成一个Map，相当于解除了我们对于高精度地图的高度依赖感，同时可以理解每个时刻的不同物体，相对于地面和道路这些位置，并且可以做之后的预测，相当于之后的路径规划问题。



我们采用DDPG算法改进自动驾驶决策的部分，同时加入礼让的驾驶概念，就是我们在遇到问题的时候要首先想到先做避让，也就是主动避让的情况，连续动态的情况下可以让自动驾驶汽车避免发生碰撞。

$$P(N_i|x) = \binom{n}{N_{i,1}, N_{i,2}, \dots, N_{i,k}} \prod_{j=1}^k x_j^{N_{i,j}}$$

$$f_i(x) = \frac{\sum_j P(N_j|x) U_{j,i}}{1 - (1 - x_i)^k}$$

那么“礼让”这一词最早起源于机器人，但机器人的速度会比较慢，如果转移到车辆方面其实还是有些难度的，而且高速当中的礼让应该还是比较困难的问题，所以这也是我们日后工作的难点。决策方面我们可能会结合PPO与HER的思想，个人比较喜欢这两种算法，所以会结合在里面，自动驾驶在高速运行的情况下也会需要一个快速决策的过程，所以选用PPO算法使得速度能够提升。

驾驶一段时间以后我们会在第二次自动驾驶的时候总结第一次的经验，因为人都是在经验当中不断积累，日后才能达到会开车的水平，所以我们也在说学习驾车的思想，然后通过HER促进自动驾驶车辆，总结之前的经验，使其在之后的驾驶过程当中少犯错误，尽量避免发生不必要的危险。决策的过程当中个人还是比较喜欢Actor-Critic机制，通用reply buffer是我们对之前驾驶的经验及其所得到的Policy的存储过程，之后的驾驶任务当中遇到类似的问题直接可以采用这种经验，不需要再做其它的改变或者训练。

控制方面主要还是RSS模型上面做出一些改进，因为不可能只是关注到前方的避让或者碰撞，也要关注后方，别人撞你的时候应该怎么办，所以采用的是双保险的机制，为了保证自动驾驶汽车的安全。当然如果传感器检测到有危险，或者是距离太近的情况下，自动驾驶汽车会直接进入安全机制，或者是作出礼让的行为，因为我们贯穿始终的都是礼让自动驾驶。

仿真平台TORCS属于3D赛车模拟游戏，个人比较喜欢通过这个来玩赛车游戏，做的效果是很好，而且是世界通用的赛车游戏，也是相对有说服力，效果会比较好一点，但是场景单一，不适合在复杂场景下做训练。

结论与展望：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。

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。德国联邦交通和数字基础设施部委员会说过，自动驾驶系统需要更好地适应人之间的交流，也就是让车辆或者机器适应我们的生活节奏，不是我们人去适应机器应该怎么做，或者是机器人之间的交流，总体来说就是以人为主，包括之后发生不可避免事故的时候主动的决定权，包括最终行为的决定权，必须要归人来掌握，尤其是必须归驾驶员掌握。吴焦苏老师的一句话让我印象深刻：“自动驾驶系统的安全性不能得到严格保证之前不应当被批准量产”。其实这也是对我们生命的负责任，因为如果自动驾驶车辆不能保证百分之百不发生事故，或者不能保证百分之百不会撞击的话就不能上路，因为我们要对自己的生命负责，也要对他人生命负责。

精彩视频抢先看




获取完整PPT，请后台回复：学术报告

供稿人：赵盼云

北理工大数据创新学习中心
让学习成为乐趣
长按二维码，关注我们



文章转载自公众号

 读芯术 >