货运数据分析及优化

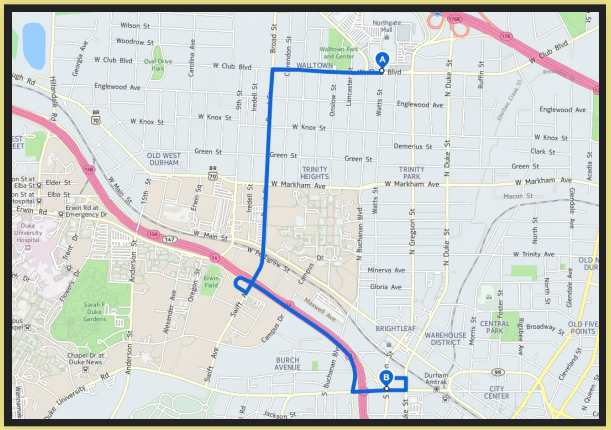
忻斌健

2020年11月26日

- 货运规划场景
- 单车动态场景
- 大数据与深度学习
  - 大数据的成功案例：深度卷积网络
- 货运优化问题难点
- 强化学习的特点
- 强化学习建模
  - MDP求解
  - Mountain Car
  - AlphaGo
  - Learning to Drive in a day (2018)
- 货运模型
- 概念验证方案
  - 求解
  - 数据接口
- 概念验证方案
- 挑战

货运规划场景

卡车 **路线**



- 单车将货物从 A --> B（旅程Trip有始有终的事件episodic, 有途经点）
- 目标：快，经济（节能），安全
- 导航规划（高速，高架，城区道路，郊区道路）
  - 道路长度
  - 高速收费
  - 交通实况
- 特点（图商提供）：
  - 与车辆状态无关（续航里程，诊断状态）
  - 按交通实况导航（不是真正的路况预测）

单车动态场景




- 其他交通参与者：其他车辆，行人
- 实时路况：拥堵，交通灯，潮汐车道
- 道路属性：长度，曲率，坡度（地面/高架）
- 实时位置，住所，服务时长，危险品属性，设备特性，驾驶员特性，货物属性

  单车 $\longrightarrow$ 十几~几十维状态向量

  环境 $\longrightarrow$ 上万维状态向量

  决策 $\longrightarrow$ 十几维决策向量
- 车辆控制：加速，刹车，转向（连续）
- 行为决策：变道，跟车，转向（离散）
- 导航决策：途径点位置顺序实时动态选择（离散）

  任务 $\longrightarrow 10^{20}$ 属性组合，（全场景）

关于动态资源分配的优化问题

大数据与深度学习

大数据的成功案例：深度卷积网络

ImageNet状态空间：

224*224 RGB图像：$2^{256^{3^{224x224}}}$ **极其稀疏！**

- CNN高效利用特征可分解，顺序致密卷积
- 高维参数的非线性函数逼近
- 利用大量实际数据逼近真实的数据分布
- 分类，检测应用，监督学习，**i.i.d.** $\longrightarrow$ 极大似然估计
- 解决了代表学习的问题：
  - 确定了深度结构 (Darknet, Transformer)
  - 提供了优化算法（随机梯度下降Adam）

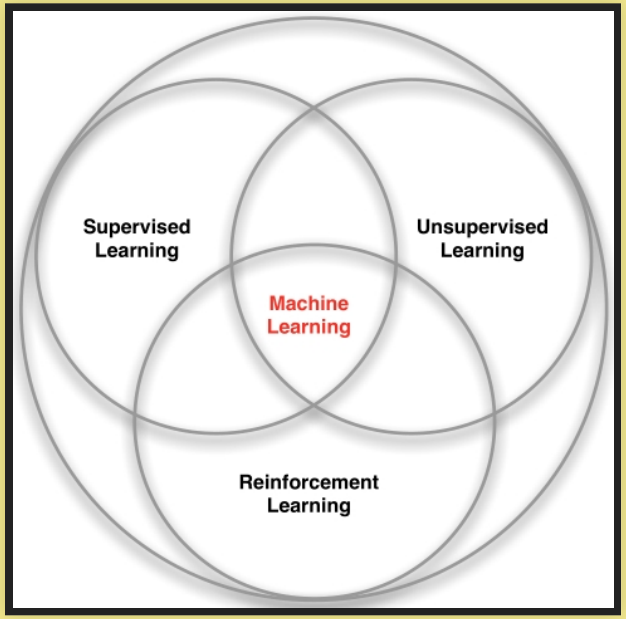
货运优化问题难点

- ***随机**过程：*
  - 状态，决策都是随机分布
  - 随机分布通常未知(如分布已知可使用动态规划求解)
- *非平稳过程：*
  - 环境非平稳，有突发事件
  - 策略优化过程导致策略非平稳
- *时间序列：非iid 分布，前后帧高度相关（数据方差大，学习收敛慢）*
- *环境部分可观测：*
  - 感知有盲区
  - 感知模式有限（雨雪，夜晚，失效/故障）
- *高维度环境状态和决策空间*

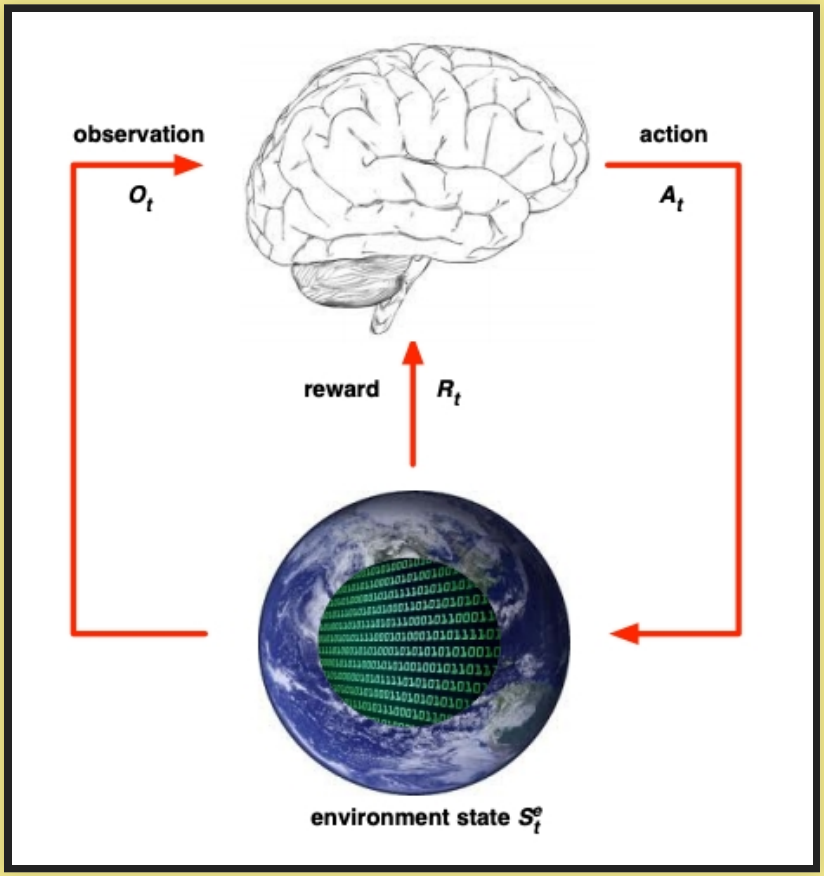
高维度环境→无法使用基于规则方法：无法建模或者建模精度不够
非稳态随机过程→不能使用模仿学习/监督学习

强化学习的特点

- 通过与环境交互学习最优策略（基于大数据）
- 积累经验（缓存在状态/行动价值函数中）
- 可以是无模型或动态建模
- 可以是在线算法
- 动态调整策略
- 适应部分可观测环境
- 适用非平稳随机过程
- 适应高维度环境，连续状态空间



强化学习建模



- Agent: 主体，智能体/控制器,**系统**
- 环境：客体，交互对象（产生观测量和奖励，接收行动）
- 信号：观测，行动，奖励

$\Rightarrow$ **SARPγ模型**

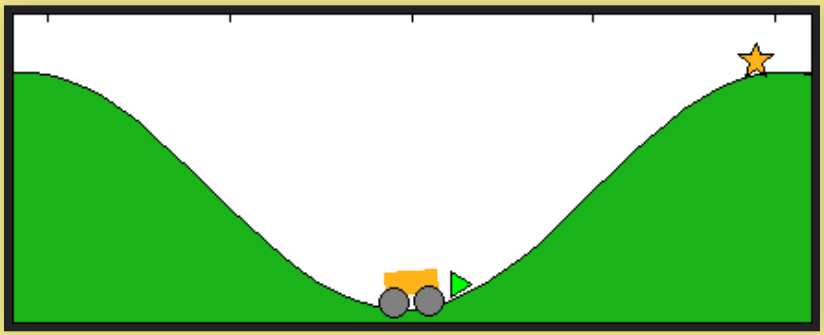
MDP求解

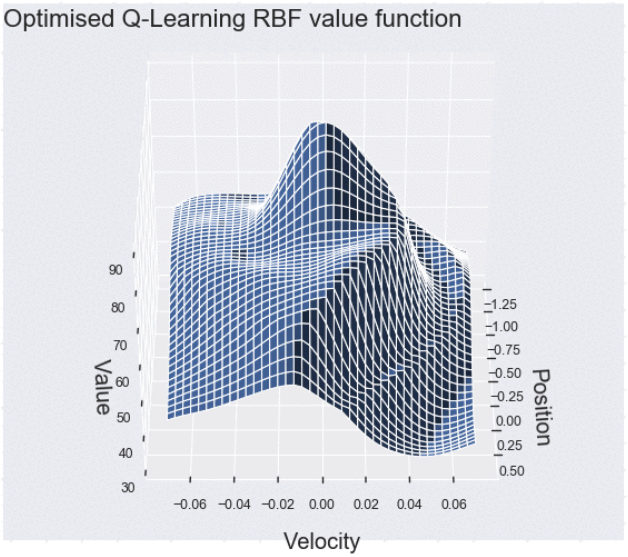
SARPγ

- State: 系统状态，对观测的最小充分描述（马尔可夫决策过程,当前的状态描述），环境状态（真实环境的全知模型）
- Action: 行动 （加速度，刹车，转向，导航决策...）
- Reward：即时奖励（到达目的地+1，用时t，能耗e；可正可负，必须是一个标量，可通过加权转换成标量）
- Probability： （Transition Probabiltiy）状态迁移概率，描述系统动态
- γ：即时奖励的时间折扣系数

> *给定环境：P R*
> *通过采样环境P和R 得到系统状态S与即时奖励R*
> *选择最优策略（Policy）： 确定性的$\pi_\theta(S)=a$或者随机的*
> $\pi_\theta(a|S)=\mathbb{P}[A_t=a|S_t=s]$
> *时间折扣系数γ是超参数*

MOUNTAIN CAR

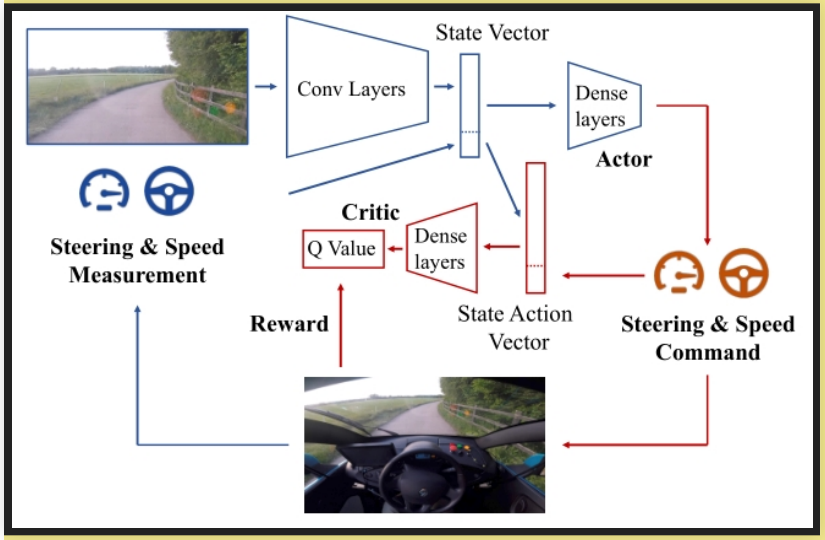




ALPHAGO

完全可观测MDP
高复杂度

LEARNING TO DRIVE IN A DAY (2018)



Learning to drive in a day

货运模型

- Agent: 主体, 系统 , **卡车**
- 环境：客体，**道路+车辆状态**
- 信号：观测，行动，奖励

$\Rightarrow$ **SARP**γ模型?

概念验证方案

求解

SARPγ

- State: 系统状态（车辆状态+道路状况）对观测的最小充分描述（马尔可夫决策过程,当前的状态描述），环境状态（真实环境的全知模型）
- Action: 行动 （加速度，刹车，转向，导航决策...），**策略参数化**（初始化可以是随机选定）
- Reward：即时奖励（到达目的地+1，用时t，能耗e；可正可负，必须是一个标量，可通过加权转换成标量）**通过采样得到即时奖励**
- Probability： （Transition Probabiltiy）状态迁移概率，描述系统动态 **通过采样得到下一个状态**
- γ：即时奖励的时间折扣系数

给定环境：P R
通过采样环境P和R 得到系统状态S与即时奖励R
选择最优策略（Policy）：确定性的$\pi_\theta(S)=a$或者随机的$\pi_\theta(a|S)=\mathbb{P}[A_t=a|S_t=s]$
时间折扣系数γ是超参数
需要很多数据

数据接口

系统状态：

- 自车状态：车速，位置，航向，诊断信号
- 道路状态：曲率，高程，俯仰
- 动态目标状态：其他车辆，行人

决策:

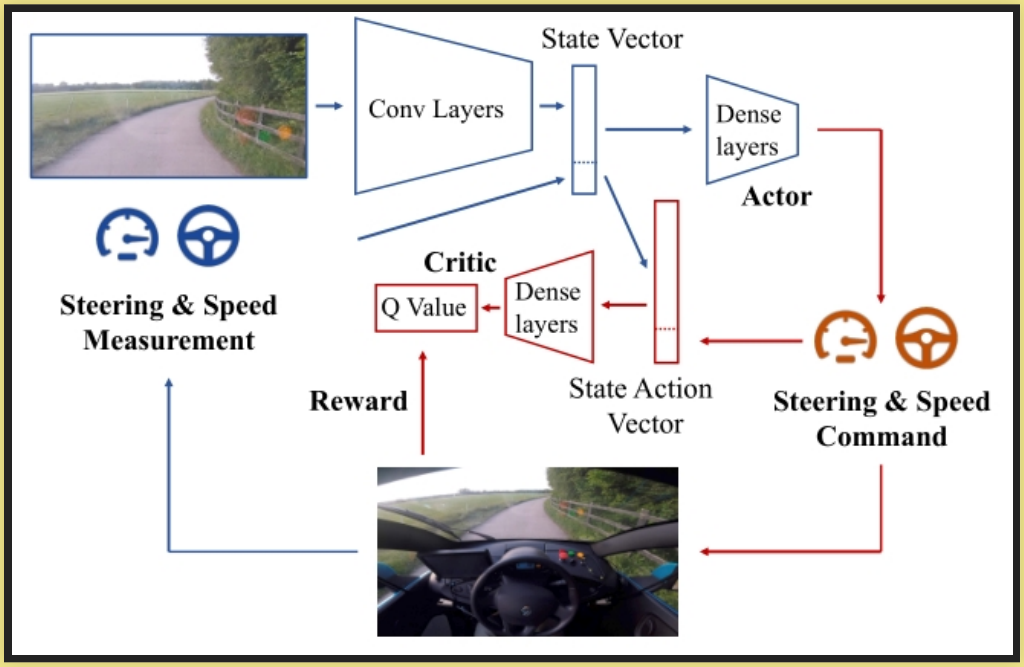
- 控制：加速度，刹车，航向 (连续)
- 导航决策：
  - 结合车辆状态的路线选择
  - 途径点选择

单一即时奖励：

- 能耗
- 成本 （能耗+时间成本+）
- 体验 （急刹少，平稳）

ETA预测数据要准确（-->图商）

概念验证方案



- 替换奖励为能耗
- 道路+车辆状态, 编码
- 道路，城市，天气，相对稳定

挑战

- 高维连续状态特征：选择和压缩?
  - 特征状态的动态调整?
- 人机结合策略?
- 部分可观测过程（POMDP)：有些关键特征不可观测
- 不违反安全约束：风险/安全-->价值函数评估
- 可解释性
- 动作：只控制纵向加速度/横纵向联合控制
- 连续+离散输入信号是否可以同时编码
  - 连续与离散决策分开
- 离线训练
  - 样本不够
  - 在统计意义上评价系统也需要较多的数据
- 多目标的奖励系统