Applications of Deep Reinforcement Learning in Communications and Networking: A Survey

NETWORK ACCESS AND RATE CONTROL

物联网等现代网络在本质上变得更加去中心化和ad-hoc。在这样的网络中，传感器和移动用户等实体需要做出独立的决策，信道和基站的选择，以实现自己的目标，如吞吐量最大化。

但网络状态具有动态性和不确定性。

- *Dynamic spectrum access:* 动态频谱访问允许用户在本地选择信道，以最大限度地提高其吞吐量。但用户可能没有对系统的完整观察，so use DQL
- *Joint user association and spectrum access:* User association is implemented to determine which user to be assigned to which Base Station. (joint user association and spectrum access problems in[42][43]) 这是个非凸优化组合问题，so use DQL（提供分布式的解决方案）

> 凸函数的局部最优解就是全局最优解，在数学中的一个非凸的最优化问题也就意味着局部最优解并不是全局最优解,所以非凸函数的寻优是最难的
>
> 因为非凸，所以要对全局都要有一个了解 。即需要接近完全和准确的网络信息来获得最优策略

- *Adaptive rate control:* HTTP上的动态自适应流(DASH)系统，其允许客户端或用户独立选择不同比特率的视频片段下载。目标就是最大化其体验质量(QoE)。so use DQL

> 不用动态规划的原因：动态规划的复杂性高，且需要完整的网络信息。

Network Access 网络接入(spectrum access & user association)

> i.e. 拉丁语的id est 意为"that is"即 e.g. 拉丁语的exempligratia 举例

channel access

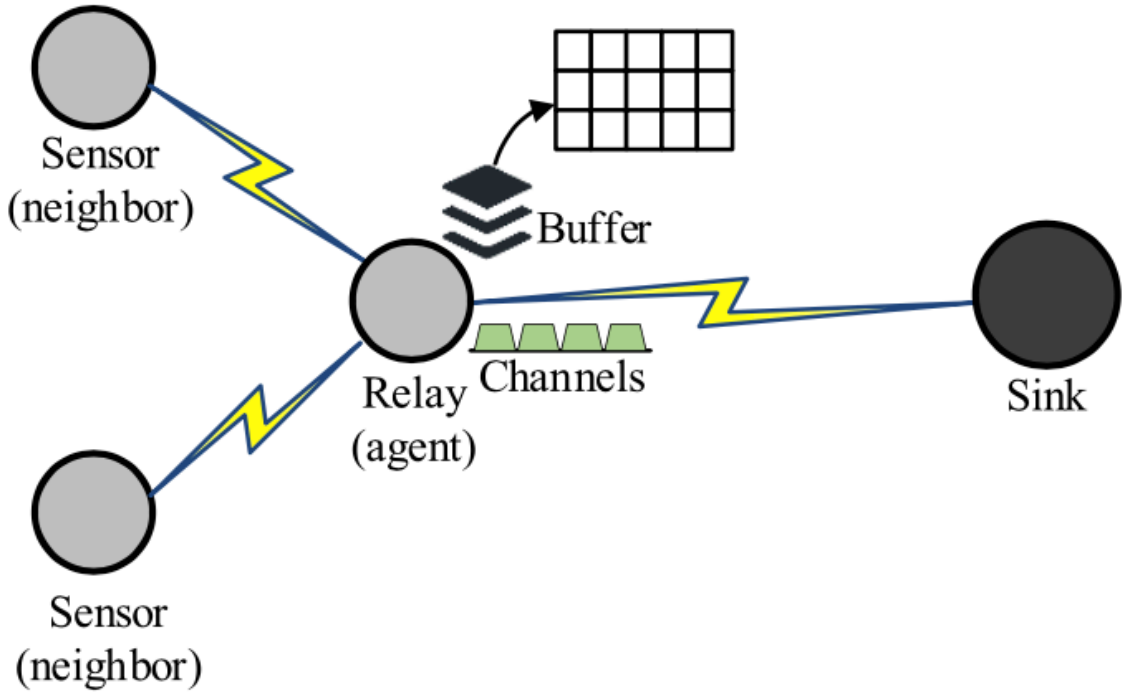
- 传感器选择M条通道来传输网络包，根据传输后的反馈，好链路 reward "+1“ 不好的链路 reward"-1".
  - 目的：找到一个最优的策略来最大化sensor’s expected accumulated discounted reward
  - 物体之前是选择短视(myopic)策略这个方案 但myopic策略需要知道system transition matrix
  - 现在用 DQN 的经验重放(experience replay)策略
  - DQN 输入state(action & reward) 输出Q-values(action相关的Q-values) adopt ε-greedy policy
  - 结果：该方案的平均奖励值为4.4， 接近于myopic策略的4.5。

DQL keeps following the learned policy over time slots and stops learning a suitable policy. But IoT environments are dynamic, the DQN in the DQL needs to be re-trained

- adaptive DQL scheme is proposed 该方案评估当前策略每一时期的累积奖励，当reward低于给定的阀值时，DQN会被重新训练去找到一个new good policy。

上面的都是one sensor，现在考虑multi-sensor的场景

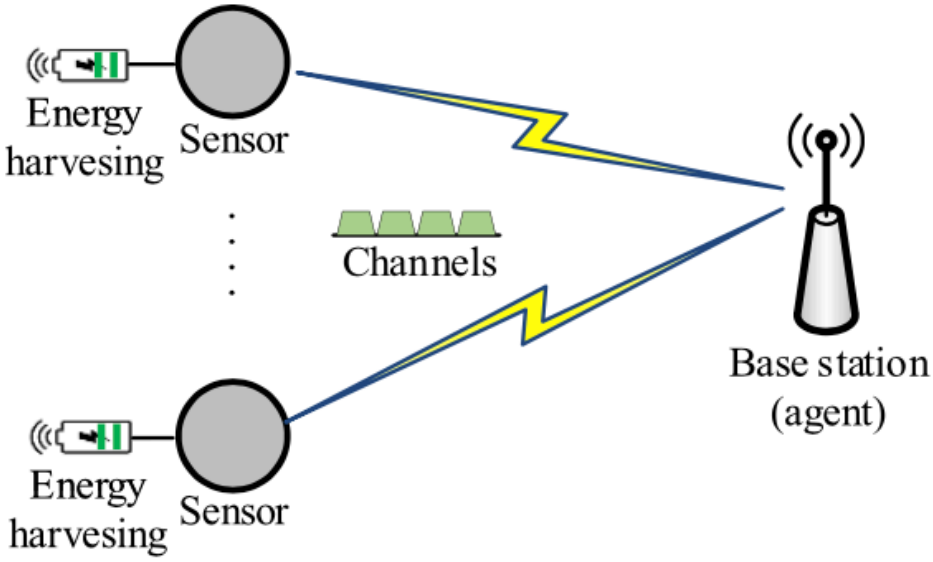
> joint channel selection and packet forwarding 联合通道选择和包转发



一个传感器作为中继(relay)，将从相邻传感器接收到的数据包转发到接收器(sink)。Relay（中继节点）有一个buffer，来存储接收到的数据包。在每个时间序列中，这个sensor选择一组通道(能最大化 (发送数据包的数量 : 发送功率))来转发数据包。

- the sensor's problem can be formulated as an MDP
    - action: 选择一组通道 通道上传输的数据包数量 和 调制模式
    - state： 结合buffer state 和 channel state
    - 输入是 state 输出是要选择的action
    - 传感器的效用函数是有界的，所以算法被证明是收敛的。
    - 结果：与random action selection scheme相比，该方案显著提高了系统的效用。
    - 不足：随着数据包到达率的增加，由于传感器需要消耗更多的功率来传输所有数据包，因此该方案的系统效用会降低。

The channel access problem in the energy harvesting-enabled IoT system




```
* BS作为控制器为传感器分配信道。
* 然而，由于传感器能量可用性的不确定性，就可能使信道分配效率低下。比如：给那些能量不多的传感器分配信道是不划算的，因为他们很快就不能用了
```

- BS的问题是：预测传感器的🔋状态，并为channel access选择传感器，以使total rate最大化

```
* 过去：使用上行资源分配方案      缺点：该方案要求BS对所有随机过程都有 perfect non-causal knowledge。
* 但是传感器随机分布在一个地理区域内，所以可能无法获得 perfect non-causal knowledge。 so use DQL
* DQL使用由两个基于LSTM的神经网络层组成的DQN。第一层来预测传感器的电池状态，第二层利用预测的状态和通道状态信息(CSI)来确定通道访问策略。
* state集合包括：(1)通道访问的分配历史； (2) 预测的电量信息历史； (3)真实的电量信息历史;(4)传感器当前(Channel State Information)CSI。
* action集合包含：被选择过的传感器集合
* reward是：总速率和预测误差之间的差值。
* 结果：该方案在总速率上接近最优方法[52]，优于myopic策略[45]。此外，该方案获得的电池电量预测误差接近于零。
```

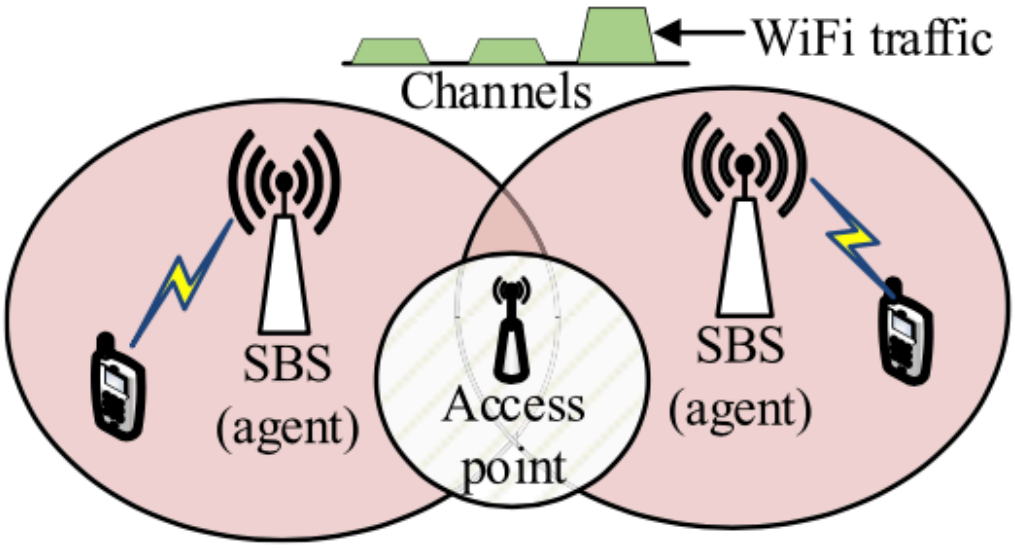
以上的方法都关注优化rate maximization 但在V2V系统中，延迟也要考虑

- 每个V2V transmitter/receiver面临的问题：在约束延迟时间的情况下选择信道和发射功率，使其容量最大化。
- DQN中each V2V transmitter的action：选择信道和选择发送功率
- reward是：有关V2V transmitter容量和延迟的函数
- state包括：(1)对应V2V链路的瞬时CSI (2)前一个时隙中 V2V链路的干扰 (3)在前一时间段内，V2V发射机的邻居所选择的信道 (4)满足延迟约束的剩余时间。
- 输入：state action 输出：该action所得到的Q-values
- 结果：在车辆链路有可能违反延迟约束时，来动态调整功率和信道选择。该方案对比随机信道分配方案，满足延迟约束的车辆发射机数量更多了。

为了降低频谱成本，上述IoT系统通常使用未授权(unlicensed)的信道。但这可能对现有的网络产生干扰。

> 什么叫 unlicensed channel?

(这个应用不太懂)利用DQN将动态信道接入和干扰管理问题一起都解决：



Unlicensed channel access in LTE networks.

> SBS(Small Base Station). LTE network: Long-Term Evolution(一个标准) network

在每个时隙，SBS选择一个通道来传输其数据包。但是，所选通道上可能有WLAN通信，因此SBS概率性地访问所选通道。

- SBS的action：信道选择和概率性地访问信道
- SBS的问题是：确定一个action vector，以便在所有通道和时间段内最大限度地提高其总吞吐量，即最大化效用函数。
- **（这里为什么又谈及资源分配）**资源分配问题可以表述为一个非合作博弈（non-cooperative game），利用基于LSTM的DQN可以求解该博弈。
- DQN输入：该通道上SBSs和WLAN的历史流量　输出： SBSs的预测action vector

The utility function of each SBS is proved to be convex, and thus the DQN-based algorithm converges to a Nash equilibrium of the game.

> 证明了每个SBS的效用函数是凸的，因此基于DQN的算法收敛于博弈的纳什均衡。

- 结果：与标准Q-learning相比，该方案的平均吞吐量提高了28%。

(这个也不是很懂)多用户共享K个信道的动态频谱访问问题

在某个时隙，用户以一定的尝试概率选择信道或选择根本不传输数据包。

1) state: 用户历史的action和当前的obeservation

2) 用户的策略是: mapping from the history to an attempt probability.

3) 问题：找到一个策略向量，也就是policy, 从而maximize its expected accumulated discounted data rate of the user

以上的问题训练一个DQN来解决

- 输入：past actions 和 the corresponding observations. 输出：estimated Q-values of the actions
- 为了避免Q-learning的过高估计，我们使用DDQN来解决这个问题。

the multichannel random access is modeled as a non-cooperative game, the game has a subgame perfect equilibrium.

Note that some users can keep increasing their attempt probability to increase their rates. This makes the equilibrium point inefficient, and thus the strategy space of the users is restricted to avoid the situation.

用户的这种策略空间(不断增加尝试的可能性，以提高其成功率)是被限制的

- 结果：该方案的信道吞吐量是slotted-Aloha [56]的两倍。原因是，在该方案中，每个用户仅从其局部观察中学习，没有在线协调或载波感知

在上述模型中，用户数量在所有时间段都是固定的，不考虑新用户的到来。

该系统的问题是找到一种信道分配决策，使新UT(User Terminals)在时间段内的总服务阻塞概率最小，同时又不会对当前UT造成干扰。

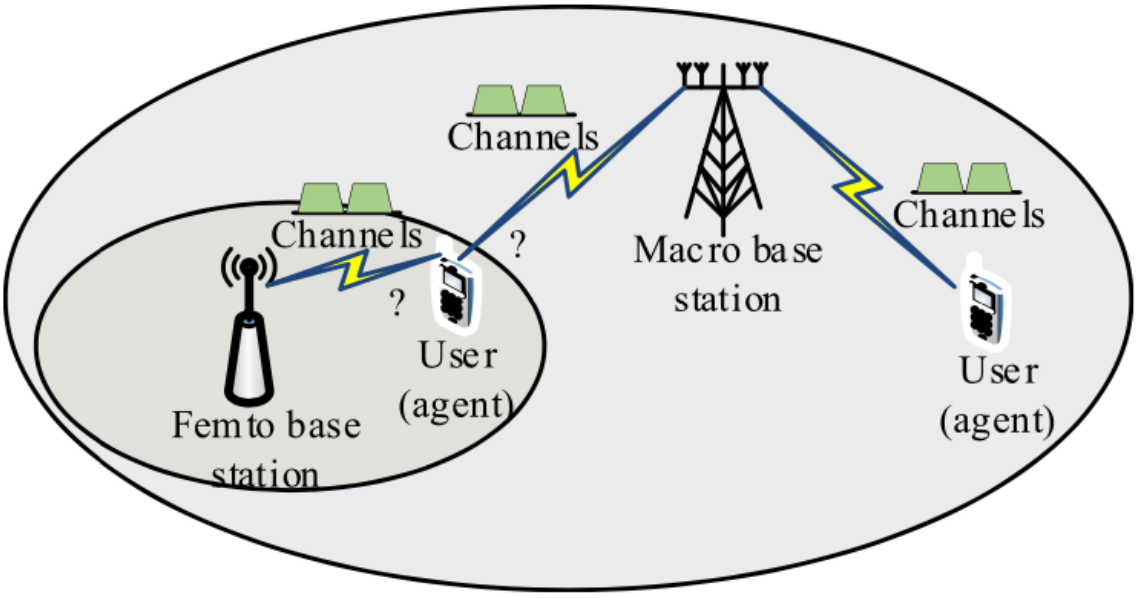
The system's problem can be viewed as a temporal correlated sequential decision-making optimization problem.

- Agent: satellite system
- Action： is an index indicating which channel is allocated to the new arrived UT.
- state集合: current UTs, the current channel allocation matrix, and the new arrived UT. (由于同信道干扰，状态具有空间相关特征,所以可以用image tensor来表示。因此，DQN采用CNN来提取状态的有用特征)
- reward: is positive when the new service is satisfied and is negative when the service is blocked
- 结果：通过将可用信道分配给新到达的UTs，与固定信道分配方案相比，该方案可以将系统流量提高24.4%。
- 不足：随着UTs数目的增加，可用通道数目很低，甚至为零。此时，所提方案的动态信道分配决策变得毫无意义，两种方案之间的性能差异变得不显著。在未来的工作中，可以研究一种基于DQL的信道和功率联合分配算法。(a joint channel and power allocation algorithm based on the DQL can be investigated.)

Joint User Association and Spectrum Access

- Joint user association(多用户归属) and spectrum access problems 是典型的非凸优化问题

> 以前采用了线性规划等传统方法来获得最优解。但这些方法几乎需要知道完整并且准确的网络信息，而这通常无法达到的。
>
> 所以使用Q-learning算法。然而，由于joint optimization problem存在较大的state空间和action空间，因此获得最优解具有很大的挑战性
>
> 所以用DQN



- 在上图中，每个用户的问题是，在保证用户的信噪比(SINR)高于最低服务质量(QoS)要求的同时，选择一个BS和一个channel，使其数据速率最大化。
- agent：每个user　　state：a vector including QoS states of all users(用户的QoS状态是指其SINR是否超过最小QoS要求)
- 在每个slot中，用户采取一个action，就会得到一个negative reward或是 positive reward, 由于一个用户的累计奖励还取决于别的用户，所以该问题可被看作MDP
- 也是使用DDQN和Dueling DQN来解决问题
- 实验表明：DQN可以有效地用于解决诸如HetNets和物联网等大规模系统中的联合优化问题

> HetNets: Heterogeneous network 异质网络：这个系统由不同操作系统组成。

use the DQL for a joint user association, spectrum access, and content caching problem.

> UAV: Unmanned aerial vehicle 无人机

- 该网络模型是一个LTE网络，由为地面用户服务的UAV组成。无人机装备有存储单元，并可以作为缓存启用LTE-BSs。无人机可以访问网络中的许可波段和非许可波段。无人机由一个基于云的服务器控制，从云到无人机的传输通过使用许可的蜂窝频段实现。
- UAV的要解决的问题为：1) 最优用户关联 2)许可频带上的带宽分配指标 3）未授权波段的时段指标 4）确定a set of popular content（用户请求它可以最大化 **稳定队列中用户的数量** 即满足内容传输延迟的用户数量）

- 无人机的问题是组合的，非凸问题
- 无人机不知道用户的请求，因此使用 Liquid State Machine approach (LSM)来预测用户的内容请求分布并进行资源分布
- UAV作为agent，使用基于LSM的学习算法来寻找最优users association
- 输入：action（other UAVs 采取的 UAV-user association schemes）

  输出：the expected numbers of users with stable queues corresponding to actions that the UAV can take.
- 用户关联完成之后，根据[61]的结果确定最优的内容缓存，并使用线性规划进行最优频谱分配。基于Gordon定理[62]，证明了所提出的DQL以概率1收敛
- 结果：DQL可以在400次迭代内收敛。与Q-learning算法相比，所提出的DQN算法的收敛时间提高了33%。与无缓存的Q-learning相比，所提出的DQL显著提高了具有稳定队列的用户数量，最高可达50%。事实上，能源效率对无人机也很重要，因此将DQL应用于联合用户关联、频谱接入和功率分配问题需要研究。

Adaptive Rate Control

- Video streaming目前主要的标准是:DASH(Dynamic Adaptive Streaming over HTTP)
- DASH能够利用现有的内容delivery网络基础设施，并与多种客户端应用程序兼容。

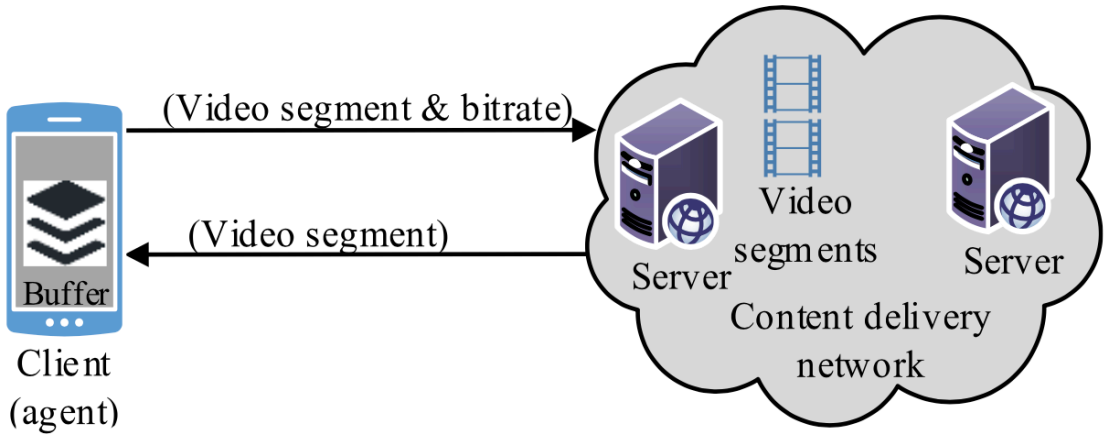


Fig. 13. A dynamic adaptive streaming system based on HTTP standard.

- 视频在服务器上存储为多个段，即块。每个片段以不同的压缩级别进行编码，以生成具有不同比特率的*representations* ，即不同的视频视觉质量。
- 在每一个time slot，客户选择一个representation(即那些具有一定比特率的段)去下载
- client问题是: 找到一个最佳策略来最大化QoE(即：最大化平均比特率和最小化rebuffering，即视频播放冻结的时间)
- 上述问题也可以建模成一个MDP： agent: client action: 选择一个representation去下载

  reward：被定义为一个函数(参数包括: visual quality of the video / video quality stability / rebuffering event / buffer state)

  State: (i) the video quality of the last downloaded segment, (ii) the current buffer state, (iii) the rebuffering time. (iv) the channel capacities experienced during downloading of segments in the past time slots.(过去所有时间)

- MDP可以通过使用动态规划来解决，但随着问题规模的增加，计算复杂度迅速变得难以管理。所以用DQL，这里使用的是LSTM network(并应用了peephole). 输入:state. 输出:Q-values corresponding to the client's possible actions
- 结果：该DQL算法比Q-learning收敛速度更快。该算法能够提高了视频质量，减少了延迟，因为其考虑缓冲区状态和信道容量来对缓冲区进行动态管理。

A3C(Asynchronous Advantage Actor-Critic)的方法提出增强和加快了训练的速度

- A3C包括两个neural network，The actor network is to choose bitrates for the client, and the critic network helps train the actor network
- actor网络中：输入: client's state 输出: policy(即：client可能采取action的概率分布。) action：choose the next representation
- critic网络中: 输入: client's state 输出: the expected total reward when following the policy obtained from the actor network
- 结果: 与bitrate control scheme相比，所提出的DQL可以将平均QoE提高25%。此外，由于有足够的缓冲区来处理网络吞吐量的波动，与baseline scheme相比，该DQL减少了约32.8%的rebuffering。

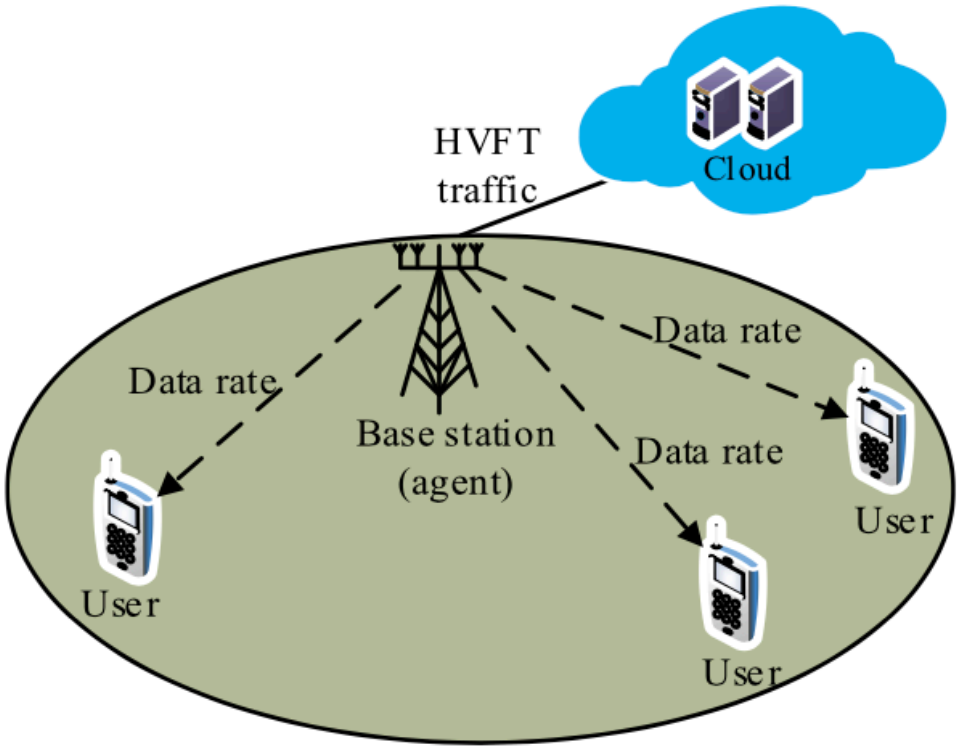
由于A3C能够支持多个agent并行训练，所以DQL易部署在多客户端网络中

- client作为agent来观察reward，其先给server发送一个tuple(state, action, and reward)
- server使用actor-critic算法来更新其actor network model，然后将这个新的model传给agent
- 这种更新过程可以在所有代理之间异步进行，提高了训练的质量和速度。
- 虽然并行训练方案在客户端和服务器之间可能会产生一个往返时间(RTT)，但[67]中的仿真结果表明，客户端和服务器之间的往返时间(RTT)仅使平均QoE降低3.5%。因此所提出的DQL可以在真实的网络系统中实现。

prediction network

- 上述输入中的client state包括：video quality of the last downloaded video segment。由于这个video segment is raw，其可能导致 "state explosion"。为了reduce the state space and to improve the QoE，所以提出了 prediction network
- 预测网络使用CNN和RNN从原始视频片段中提取有用的特征。然后，将预测网络的输出作为DQL的输入之一。
- 结果：本文提出的DQL可以将平均QoE提高25%。此外，由于DQL的状态空间小，使得视频传输的平均延迟降低了45%左右。这意味着在状态空间较大的情况下，应该使用CNN来提高用户的QoE和收敛时间

除了DASH系统，DQL还可以有效的用于HVFT(High Volume Flexible Time)的速率控制



g. 14. Data rate control based on DQL in HVFT applications.

> traffic: 流量

- 由于HVFT的应用具有较大的流量，因此需要进行流量调度，如数据速率控制
- 过去方法：为每个流量类型分配静态优先级，然后基于其优先级进行流量调度。但这种方法并不会进化。从而适应新的traffic classes。所以使用DQL来提供 adaptive rate control mechanism
- BS的问题: 找到一个合适的策略，即用户的数据速率，以最大限度地增加HVFT传输流量，同时最大限度地降低现有数据流量的性能退化。 因此这个问题也可以被建模成MDP问题
- Agent: BS state: the current network state and the useful features extracted from network states in the past time slots

  在一个时隙的state包括：cell's traffic load(小区在该时段的流量负载) 总的网络连接数 电池质量

  action包括：a combination of the traffic rate for the users

  reward被定义为一个函数：参数包括 (i) the sum of HVFT traffic
  (ii) traffic loss to existing applications due to the presence of the HVFT traffic
  (iii) the amount of bytes served below desired minimum throughput.
- **DQL 使用的也是 the actor and critic networks with LSTM**
- 结果：通过使用墨尔本采集的真实网络数据，仿真结果表明，与启发式控制方案相比，所提出的DQL方案使HVFT流量增加了2倍。因此，文中提出的DQL有望应用于人口增长较大的大型城市的现代网络。

DQL can be used for the rate control to achieve multiple objectives in complex communication systems.

- 在系统中，发射机需要配置多个传输参数，如符号率、编码率等，以实现多个冲突目标，如低误码率、提高吞吐量、功率和频谱效率等。可以使用自适应编码和调制方案，但该方法只能实现有限的目标。so use DQL
- Agent: 系统中的transmitter　Action: 是一串集合,包括(i) symbol rate, (ii) energy per symbol, (iii) modulation mode, (iv) number of bits per symbol (v) encoding rate.
- 目标是最大化系统性能。reward: 适合度函数的性能参数,包括(i) BER estimated at the receiver, (ii) throughput, (iii) spec- tral efficiency, (iv) power consumption, (v) transmit power efficiency.
- State: system performance, e.g. reward

To achieve multiple objectives, the DQL is implemented by using a set of multiple neural networks in parallel.

- 输入：DQL当前状态和信道条件　输出：predicted action 采用Levenberg-Marquardt反向传播算法对神经网络进行训练
- 结果：所提出的DQL可以达到比较理想的适合度评分，即不同目标的加权和。也是穷举搜索方法

summary

DQL在动态网络访问和自适应速率控制中的应用大部分都被建模成MDP。此外，用于 IoT (物联网)和DASH(Dynamic Adaptive Streaming over HTTP)系统的DQL方法比其他网络受到更多关注。

未来的网络，如5G网络，涉及多个网络实体，它们有多个相互冲突的目标，如：供应商的收入与用户的效用最大化。这对传统的资源管理机制提出了一些挑战，值得深入研究。

2. DQN在 CACHING AND OFFLOADING 卸载和缓存上的应用

3. DQN在NETWORK SECURITY AND CONNECTIVITY PRESERVATION 网络安全和连接保存 上的应用

4. 其它各种各样的问题

- *Traffic Engineering and Routing*
- *Resource Sharing and Scheduling*
- *Power Control and Data Collection*
- *Direction-of-Arrival (DoA) Estimation*
- *Signal Detection*
- *User Association and Load Balancing*
- *User Localization*
- *Access Device Detection*

5. CHALLENGES, OPEN ISSUES, AND FUTURE RESEARCH DIRECTIONS 挑战、开放问题及未来研究方向

挑战

- *State Determination in Density Networks:*
  - the DRL approaches often require the users to report their local states at every time slot
  - 为了观察本地状态，用户需要监控来自相邻BSs的RSSIs(Received Signal Strength Indicators) 然后暂时使用最大RSSIs连接BS。但是未来基站的RSSI将趋于相同，所以就很难选择这个临时BS
- *Knowledge of Jammers' Channel Information*
  - [126]提出的无线安全DRL方法使无人机能够找到最优的传输功率水平，使无人机和BS的安全能力最大化。然而，为了制定无人机的奖励，需要对干扰机的信道信息有一个完整的了解。这在实践中具有挑战性，甚至是不可能的。
- *Multi-Agent DRL in Dynamic HetNets*
  - 现有的大部分工作都集中在基于本地观察或交换的网络信息为单个网络实体定制DRL框架上。希望网络环境是相对静态的，以保证学习结果的收敛和政策的稳定。在动态异构5G网络中，这一要求可能会受到挑战，该网络由分层嵌套的物联网设备/网络组成，具有快速变化的业务需求和网络条件。在这种情况下，单个实体的DQL代理必须是轻量级的，并且能够灵活地适应网络条件的变化。这意味着在学习中状态和行动空间的减少，然而这可能会损害收敛策略的性能。多智能体之间的相互作用也使网络环境复杂化，导致状态空间的大幅度增加，这不可避免地降低了学习算法的速度。
- *Training and Performance Evaluation of DRL Framework:*
  - 在无线通信中，训练和测试的数据很难获取。许多已有的工作是基于模拟数据的，这比真实数据就简单很多。

open issues

- *Distributed DRL Framework in Wireless Networks*
  - DRL框架需要大量的DNNs训练。对于limited capabilities的大量终端用户来说，为DRL框架设计分布式实现是一项有意义的任务。
- *Balance Between Information Quality and Learning Performance*
  * 如何在信息质量和学习性能之间找到最优平衡点，使DQL代理不会消耗过多的资源而只获得学习性能不显著的边际增长，是一个需要解决的问题。

未来研究方向

- *DRL for Channel Estimation in Wireless Systems*

```
* Wireless-Powered Crowd Sensing (WPCS) 技术将更有前景。所以，提高WPT的功率传输效率是实现WPT在低功耗广域网中部署的关键。
* 由于传感器必须通过专用能源的WPT进行自我供电，所以基于传感器节点的接收功率，将其作为DQL的输入，从而实现对信道的估计
```

- *DRL for Crowdsensing Service Optimization:*
  - 在MCS(Mobile Crowd Sensing 移动群智感知)中，移动用户向众测服务提供商提供传感

数据，并获得奖励。然而，由于有限的资源，例如带宽和能量，移动用户必须决定是否和多少数据上传到提供商。同样地，以利润最大化为目标的提供者必须确定给予奖励的数量。提供商的决定取决于移动用户的操作。例如，由于许多移动用户向众测服务提供商提供数据，该提供商可以降低激励。由于用户数量多，状态空间大，环境动态，采用DRL可以得到类似于[194]的最优群智感知策略。

- *DRL for Cryptocurrency Management in Wireless Networks*、
  - 用于无线资源访问和服务使用，或兑换成真实的货币。在随机加密货币市场环境下，DRL可以像[198]一样，为无线用户实现加密货币管理的最大长期回报。
- *DRL for Auction(拍卖)*

> * Auction已被有效地用于无线电资源管理，例如频谱分配[199]。然而，当竞拍者(即竞拍者和竞拍者)非常多时，如何确定拍卖的结果就变得复杂而棘手了。这种情况在下一代无线网络中是典型的，如5G高密度异构网络。DRL似乎是解决不同类型的拍卖的有效方法，如在[200]。