Unravel the Secrets of Deep Learning

Intelligent User Association via Unsupervised Learning, Supervised learning, and Deep Reinforcement Learning

Hao Jiang

Nankai University
hao_jiang2019@163.com

2022 年 8 月 10 日

Presentation Overview

1. 综述
   - 工作动机与创新点
   - 系统模型与问题的提出
2. 问题的解法
   - 解法一：Unsupervised Learning
   - 解法二：Supervised Learning
   - 解法三：Deep Reinforcement Learning
3. 仿真结果
   - 仿真结果及分析

工作动机与创新点

Motivations

在 simbiotic radio 的文献中，绝大多数文献将研究重点集中在对于 BD 反射系数的优化来实现 BD 信息传输的时效性 (timeliness)，吞吐量 (throughput)，或者可靠性 (reliability) 等等。无一例外的，研究人员都假设 SR 系统中的 user association 是建立完成的，或者该环境是稀疏的 (sparsity)。因此，如何让松弛这一假设并尝试解决 user association 成为一个关键性问题。Ref.[1] 通过 DQN (Deep Q Network) 的方式解决 TDMA 中的 user association 的问题，但是其有三个方面的 limitations:

- Channel station information: 虽然文章为了松弛这一假设，尝试采用 outdated CSI 作为 DQN agent 的输入，但是在 BS 同时获得来自 BS→MU, BS→BD, BD→MU 的全部 perfect CSI 仍然是非常困难的;
- Channel correlation assumption: 文章中假定不同 frame 之间的信道信息的关联度是被 $\rho \in [0, 1]$ 所确定的，我认为这也就是为什么 DQN 在采用 outdated CSI 的前提下仍然可以获得很好的效果 (possibly，表现好仅仅因为信道信息相关度大，而非 agent 理解了信道的相关性，简单理解，一个随机生成的东西是无法被学习的);
- Scalability: DQN 在强化学习中主要解决离散输出的问题，然而当 action space 的 dimension 很大的时候，无论是 exploration 还是 exploitation 都会非常困难，这一点造成了文章中所提出方法的 scalability 是有疑问的;

Contribuations

该工作主要有以下几点的创新点

- 采用部分信道信息，即，BD→MU 的 channel gains 来实现 user association，并且不假设信道信息之间的相关性;
- 采用了 unsupervised learning 来实现有更高 scalability 的 user association method;
- 比较了 unsupervised learning，supervised learning，以及 deep reinforcement learning 的结果，通过比较证明 unsupervised learning 的优越性;

系统模型

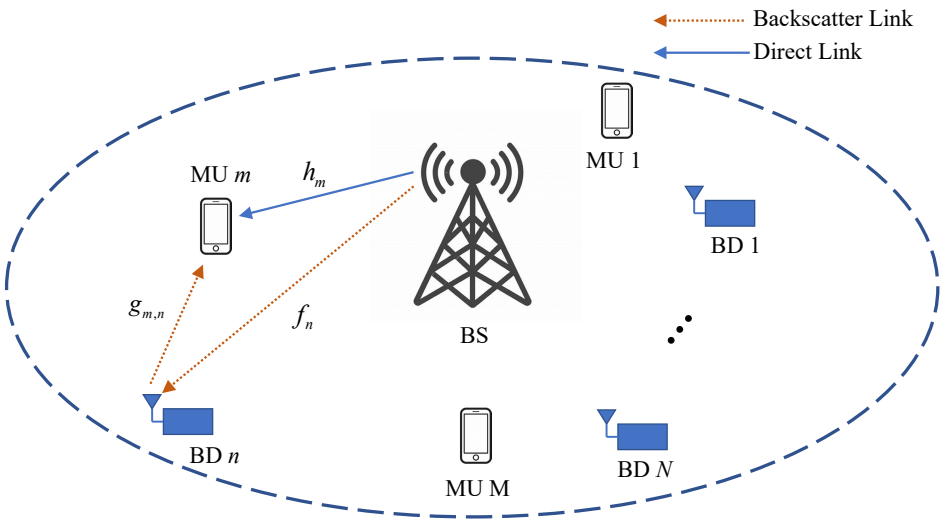


图: 系统模型

Descriptions:

1. 在 BS 的 coverage region 中有 M 个 MU 和 N 个 BD，且有 $M \geq N$，BS 通过 TDMA 的方式下行服务 MUs，同时 MU 解调于其相关联的 BD 的信息；
2. Large-scale fading model: $32.45+20\log_{10}(f)+20\log_{10}(d)-G_t-G_r(\text{in dB})$, where $f = 2.4$ GHz and $G_t = G_r = 2.5$ dB; small-scale fading coefficient: Rayleigh; The transmit power at BS is 40 dBm, the variance of the CSCG noise is set to -114 dBm, and the reflection coefficients at BDs are uniformly set as 0.8.
3. The channel matrix between BD to MU is denoted by $\mathbf{G} \in \mathbb{C}^{M\times N}$, the BS to BDs channel vector is denoted by $\mathbf{f} \in \mathbb{C}^{N\times 1}$. The channel vector from BS to MUs is given by $\mathbf{h} \in \mathbb{C}^{M\times 1}$. The binary user scheduling vector (according to TDMA protocol) is denoted by $\mathbf{t} \in \mathbb{R}^{M\times 1}$, where $\|\mathbf{t}\|_1 = 1$. The binary association vector is denoted by $\mathbf{a} \in \mathbb{R}^{N\times 1}$, where $\|\mathbf{a}\|_1 = 1$. $\mathbf{G}_{:,n} \in \mathbb{C}^{1\times N}$ is the n-th row of $\mathbf{G}$.

问题提出

1) 信号模型及解释

The received signal at the m-th MU can be given by

$$y_m = \sqrt{P}\mathbf{t}^T\mathbf{h} + \sqrt{P\alpha}\mathbf{t}^T\mathbf{G}\left(\mathbf{f}\odot\mathbf{a}\right) + v \tag{1}$$

where $v \sim \mathcal{CN}(0,\sigma^2)$ is the CSCG noise at the receiver, and $\odot$ denotes the Hardamard product.
$\mathbf{h}\in\mathbb{C}^{M\times 1}$ 是由 BS 到各个 MU 的信道系数构成的 channel vector，而 $\mathbf{t}\in\mathbb{R}^{M\times 1}$ 是 TDMA scheduler，其中只有一个 1 余下元素均为 0，代表为其中非 0 元素索引的 MU 在此时刻接收信号。项 $\mathbf{f}\odot\mathbf{a}$ 的物理含义为：$\mathbf{f}\in\mathbb{C}^{N\times 1}$ 是由 BS 到 BD 的 channel vector，而 $\mathbf{a}\in\mathbb{R}^{N\times 1}$ 为 user association vector，代表当前为其中非 0 元素索引的 BD 进行 backscatter，二者的 Hardamard 乘积代表根据 $\mathbf{a}$ 选择某一 BD 进行传输。

2) 问题提出及解释

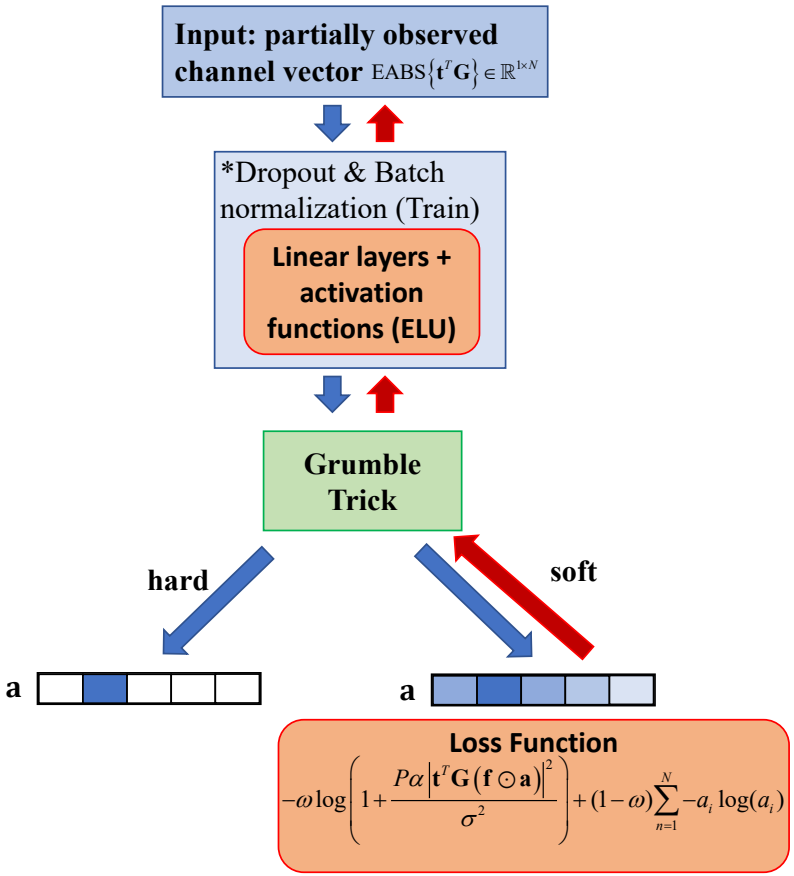
$$(P1): \quad \max_{\mathbf{a}} \log\left(1+\frac{P\alpha\left|\mathbf{t}^T\mathbf{G}\left(\mathbf{f}\odot\mathbf{a}\right)\right|^2}{\sigma^2}\right) \tag{2}$$

$$\text{s.t.} \quad |\mathbf{a}_i| \in \{0,1\} \ \forall i = 1,\ldots,N, \tag{c-1}$$

$$\|\mathbf{a}\|_1 = 1. \tag{c-2}$$

解决的问题是：在给定 TDMA scheduler $\mathbf{t}$ 的前提下，如何对于 BD scheduler $\mathbf{a}$ 进行设计来实现 throughput 最大，而限制条件 (c-1) 与 (c-2) 保证了 $\mathbf{a}$ 为 "one-hot vector"，即该向量仅有一个元素为 1 其余元素均为 0。

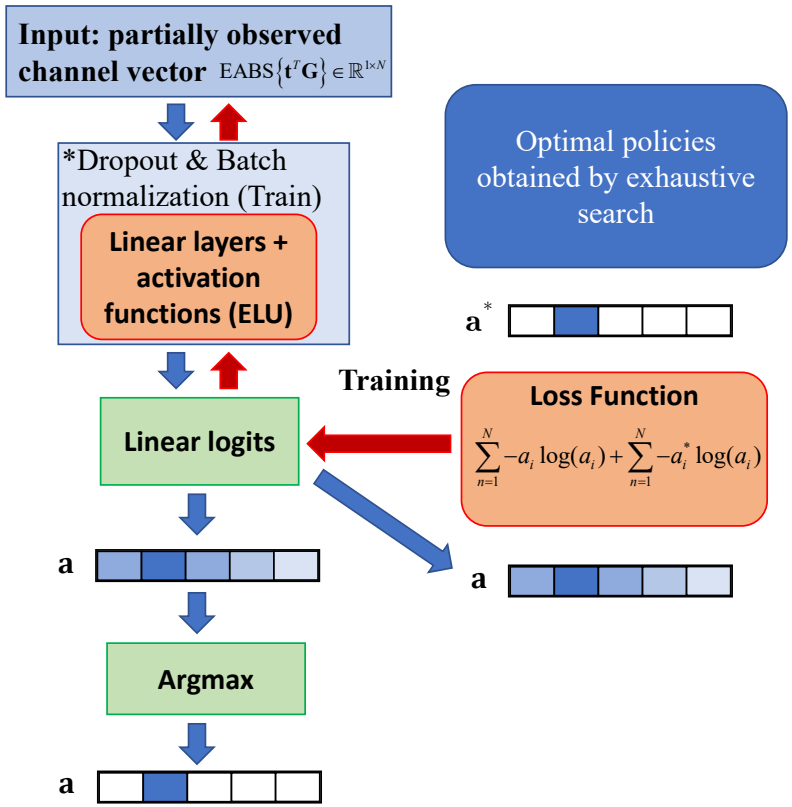
解法一： Unsupervised Learning



Descriptions:

- 红色箭头代表的梯度的流向（back propagation），而蓝色代表了数据的流向（forward）；
- 神经网络的输入是部分观察的信道信息，也就是说，只有从 BD 到 MU（根据 TDMA 协议，当前传输的 MU）的信道信息是被需要的，where $\mathrm{EABS}\{\cdot\}: \mathbb{C}^{1\times N} \mapsto \mathbb{R}^{1\times N}$ is the element-wise modulo function.
- 当神经网络处于 train 的状态时，Dropout 与 batch normalization 操作是打开的。Dropout 通过以一定概率随机暂停一部分神经元的学习，从而提升其它神经元的表达，避免 dead neural 的存在（不进行学习的神经元），防止过拟合的出现；Batch Normalization 是将一次学习的样本总和，called batch，进行 whiten 的操作（均值为 0，方差为 1）。一般认为，白化的样本是容易学习的。从通信角度讲，由于大尺度衰落的存在，样本的均值偏小，BN 的存在避免了数据预处理的麻烦；
- Grumble trick 的作用是保证梯度的正常流通。具体来说，虽然 **a** 是仅一个元素为 1 剩余元素为 0 的向量，但是在训练时我们不能这样做，因为通过 $\arg\max\{\cdot\}$ 是一个不可导的操作，因此无法允许梯度的反向传播。To this end，Grumble trick 在神经网络训练时，该模块设置为 soft 保证了梯度的传播，而当神经网络部署时，其被设置为 hard，保证了输出时的 one-hot；
- 无监督学习最重要的时对于 loss function 的设置，注意神经网络在训练时进行的是 stochastic gradient descent，所以再 loss 的第一项，即 (P1) 的目标函数，上要加入一个 minus 符号，而 loss function 的第二项代表了向量 **a** 的熵，神经网络会逐渐降低第二项的熵，进而使得 **a** 更加确定，也就是更接近一个 one-hot vector。

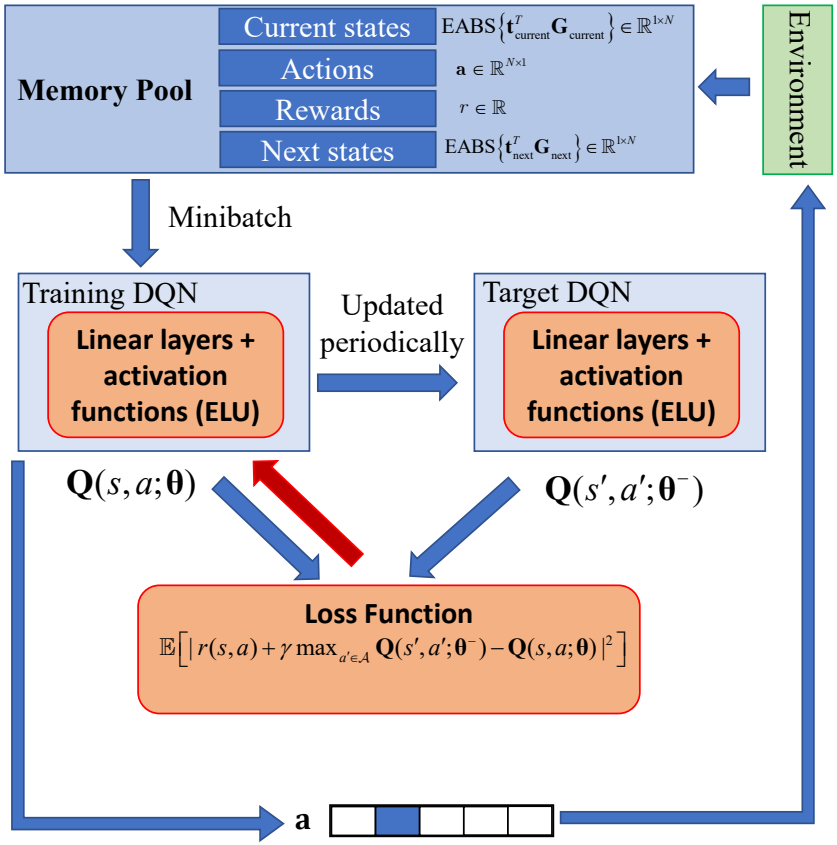
解法二：Supervised Learning



Descriptions:

- 监督学习与无监督学习的一个重要区别是前者需要首先获得训练样本，在这里我们需要事先通过 exhaustive search 的方式，获得最优的 **a** 集合，by set $\{\mathbf{a}^*\}$，这在复杂度较高的问题中常常被认为是不可行的。
- 为了公平性原则，输入与神经网络的结构都与 unsupervised learning case 保持一致，区别在于输出变为了线性输出，即没有 Grumble trick block；
- 在训练时 loss function 采用了交叉熵 corss entropy，其损失函数的物理含义同样可以分两部分理解：第一项中的为神经网络的输出的熵，同样的缩小熵是为了让其更接近 one-hot vector，而第二项为输出分布与标签分布（optimal solution via exhaustive search）的互信息量，named as KL divergence，缩小该项可以让两个分布更加地接近。

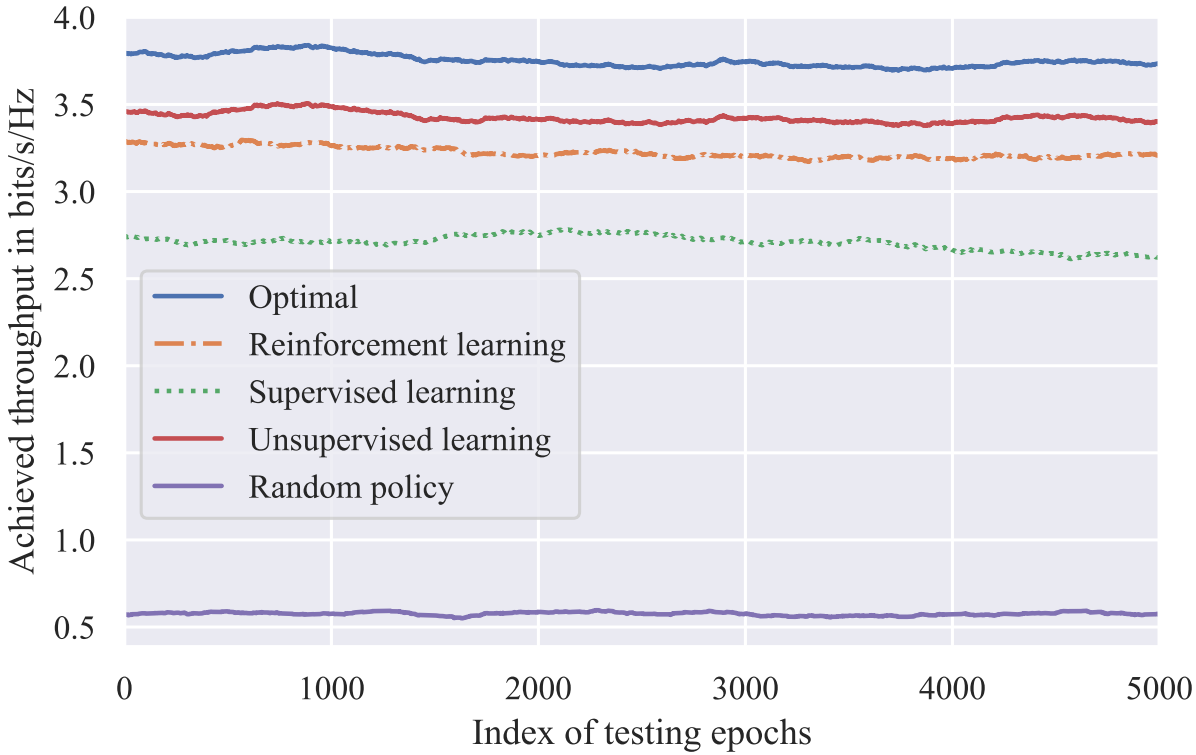
解法三：Deep Reinforcement Learning



Descriptions:

- 强化学习的框架比较简单这里就不再赘述了，值得一提的是，为了保证测试的公平性，由于 DQN 没有实际意义上的训练集和测试集（Since DQN agent learns via interaction with the environment），我的方法是首先让 agent 在与前面两种方法同样大小的训练集上进行训练，然后到同等大小的测试集上测试，但是真实条件下的 DQN 在部署时并没有这样的设置；
- 其次，在公平性的前提下，DQN agent 观察到的状态与前面两种方法的输入相同。

仿真结果及分析 I



解释

- 仿真参数的设计：BD 与 MU 的位置依照 PPP 分布生成在 $100\text{m} \times 100\text{m}$ 的范围中，BS 在地理位置上居中。大尺度衰落模型 $32.45 + 20\log_{10}(\text{f}) + 20\log_{10}(\text{d}) - \text{G}_\text{t} - \text{G}_\text{r}(\text{in dB})$, where $\text{f} = 2.4$ GHz and $\text{G}_\text{t} = \text{G}_\text{r} = 2.5$ dB，小尺度衰落模型：Rayleigh fading，反射系数:0.8，共有 10 BDs and 10 MUs，BS 的传输功率为 40 dBm，噪声功率为 -114 dBm。

仿真结果及分析 II

- 从图中可以看到 unsupervised learning 的结果最好，reinforcement Learning 的结果次之，表现最差的为 supervised learning 的结果，且它们的表现均远超 random policy，这一点表明各种网络都进行了有效的学习。
- 采用传统方法是不行的，因为在部署时没有完整的信道信息进行优化算法设计。
- Reinforcement learning：该问题并没有充分挖掘 DQN 的潜力，原因在于：1） 信道是随机的，彼此之间没有关联度，因此没有规律可循，而 DQN 解决的问题常常是 Markov decision process 类的问题，所以，DQN 被用在该情境下不合时宜；2） DQN 算法的输出是在可能的 actions 中选择 Q-val 最大的动作，但是当动作维度较大时，有一些动作的价值没有被充分挖掘（动作太多尝试的不够），导致 Agent 的决策被卡在一个 local optimal。此外，由于 DQN agent 是在与环境的互动中学习的，因此学习效率很低，不能很好的并行化（parallelization），因此训练该 DQN 所用的时间是训练 unsupervised/supervised learning network 的 60 倍。比较有趣的一点是，DQN 最大的 limitation 在于限制条件不好引入，DQN 是通过 reward 进行学习的，但是 reward 是一个 scalar，其维度很低，压缩了很多信息，比如：agent 的左右都是墙，撞上去会有惩罚（很小的 reward），但是 reward 本身并不能反映撞到了哪边的墙，agent 只能通过 action 综合去判断，而在一些 observation 与 reward 没有很强联系的情景下（比如通信中优化目标下面的限制条件），限制条件不好加入；反观 unsupervised learning 可以将限制条件乘上一个权重加入到 loss function 中，可见 unsupervised learning 对于限制条件的友好性。
- Supervised learning 的挑战主要有两个，其一是需要通过 exhaustive search 的方式寻找最优解，当输出维度很大或者问题比较复杂的时候，这样的方式是 too expensive to be practical 的，其二，从 deep learning 与 communication 结合的角度来讲，在 supervised leaning 的情况下，NN（neural network）进行将该问题看作一个存粹的分类问题，即单纯寻找最佳匹配，却忽略了输入与输出在通信上的内在联系，这样也导致了 degradation of its performance.

References

Q. Zhang, Y.-C. Liang, and H. V. Poor, "Intelligent user association for symbiotic radio networks using deep reinforcement learning," IEEE Trans. Wireless Commun., vol. 19, no. 7, pp. 4535–4548, 2020.