ADL—Deep RL

2019年4月16日　　上午 09:31

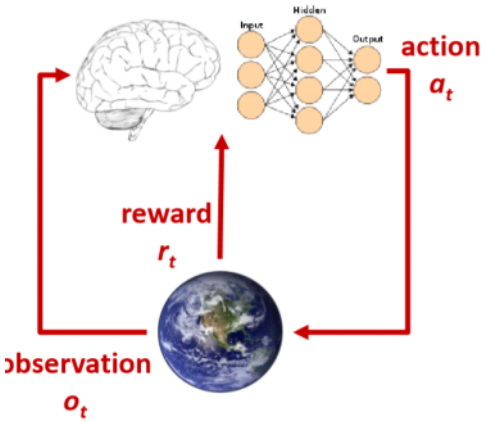
RL是基於reward的signal
依據最後是贏還是輸來決定怎麼修正
一直到全部結束才會有signal好或是不好，feedback delay
supervised learning每個時間點都會被當成獨立的來看
RL每個時間點都很重要，下在不同位置會得到對方不同下的旗，所以每一步不同得到的環境就不同
RL不知道那些環節不好，得到的只有最後互動過程後的reward (learning from critics)

- Intro
  - 一連串時間點的decision making
  - agent: 能夠做很多action 可以去選擇
  - 每個action會導致之後看到的狀況不一樣
  - 依照互動決定，希望reward越大越好

  - DL可以想成一種representation把x轉成y
  - 希望可以學到好的feature可以達到y的目標
  - 通常x是raw input(EX:pixel、signal)

  - Deep RL被認為比較接近人的行為，可以跟環境互動
  - DL給出架構，RL給出目標reward
  - RL+DL比較類似泛AI，跟現實生活比較像
  - 機械手臂如果用RL，就可以自己學手臂旋轉方式，不用明確固定寫死，依據跟環境的互動就好了
  - 跟使用者互動的也都可以用RL，做成feedback當成reward去update推薦的程序
- 概念
  - agent做action,影響state,去最佳化reward
  - agent跟env是兩個互動的對象，agent用DL就是一個model
  - 給不同action有不同state，agent就會依據看到的東西做某一個動作，獲得reward
  - feedback不一定立即
  - EX: 希望可以學到怎麼左右移動，讓分數可以最大化
  - agent看到observation做一個動作，env依據action給出下一個reward

Agent and Environment



At time step t
- The agent
  - Executes action a_t
  - Receives observation o_t
  - Receives scalar reward r_t
- The environment
  - Receives action a_t
  - Emits observation o_{t+1}
  - Emits scalar reward r_{t+1}
- t increments at env. step

- state可以被想成是一連串的observation、，根據環境的state去反映現在的狀態,state包含了從頭到尾的observation, reward,等資訊，才可以反映現在的狀態，要描述狀態要把全部記下來，所以中間過程一個不一樣，所得到的狀態就會不一樣，要記錄下來才可以知道怎麼做決定，讓下一件事情發生的資訊。state包含所有可以讓你預測下一步的所有資訊
- state可以讓你決定下一個時間點會發生事情的所有資訊，experience history去推測下一個時間點
- 有state 可以包含所有用到的資訊，可以幫助agent去選擇比較好的action
- env會用來選擇observation跟reward
- 做的所有動作跟得到的資訊跟reward放進去function
- env的state是看不到的，決定下一個時間點吐出的observation跟reward
- agent state，agent internal可以知道的資訊，決定這個agent下個時間點選擇哪個action，用RL學的資訊
- agent state: experience function
- information state: Markov state，始有資訊去判斷接下來要發生的事情
  - 給定前一個state，跟看歷史紀錄都相同，所以根據st就可以預測st+1了，因為所有資訊都會反映在st上面
  - 未來發生的事情只跟現在相關
  - 希望st就包含足夠多資訊就可以說出未來會發生的相關資訊
  - 只要知道現在時間點的state就可以不用管history了
- 不實際=>fully observable env: 如果agent看的到所有env狀況，就可以直接把env的state當成自己agent的state就可以知道該怎麼互動，當成observation就可以決定最好action=>MDP
- 真實=>partially observable env: agent看到的資訊是間接的，不可以直接反映env的情形=>POMDP (partially observable markov decision precess)
  - 對話，因為可能語音辨識錯誤，那要怎樣給出好的action

        - 去學env可能的長相，把所有可能都算一個機率分布，模擬對手可能會下的所有機率分布當成state去做學習，基於他最可能怎麼做
        - 利用NN的hidden state vector就當成給定的state，給定vector去學
        - reward
    - Reward
        - 會說現在agent做得好不好
        - 希望agent可以maxmize reward 加總的期望值
        - total reward越高越好
        - 要去maxmize 最終reward，但是action可能是非常長遠的。reward會delay，有時候可能還要犧牲當前reward，去最大化最終reward
- Agent看到observation當成agent state，要去選擇ACTION。state是要自己決定的，怎麼選一個state，互動方式會改變環境
- agent又叫做actor也叫做policy，要predict出對應action去maxmize reward
    - input: observation、ouput: function、reward調整
    - 每一個時間點多的reward才是reward，而不是每個時間點都給總分，這樣會跟最終reward不同
- alpha go ，先supervised 棋譜，再去自己跟自己下棋得reward。alpha zero，自己胡亂下，還可以學到很多不知名下法，但是所需要的時間極久。supervised + RL的結果比較好
- chatbot的學習方式，會跟reward有關
    - EX: 對話越長越好or對話越快結束越好or對話要跟主題有關
    - 對話評估好壞不容易，很抽象
    - 要盡可能maximize reward，只是不一定對化學道只是符合criteria不一定真的好
- 玩一場是一個episode，很多場的資訊可以讓model學到如何可以最大化期望值

Markov Reward Process (MRP)

- discount reward，月後面的reward discount越多

- Value function: 評估每個state、或是配上action這樣好不好
    - 價值評估
- Policy: 有一個function可以把state去map到動作上面
- Model: 要去學環境怎麼互動EX: 對手怎麼回應去評估怎麼做比較好

Value-based RL

- Q function
- 評估每個時間點state的value，往value高的走，依據值去決定怎麼做動作比較好
- 依據分數走

- 不會有明顯的policy

Policy-based RL

- pi function
- 去得到future reward 最大值
- 每一個state去學出要去做的action 怎樣比較好
- 每個state要做哪個動作
- 沒有value function

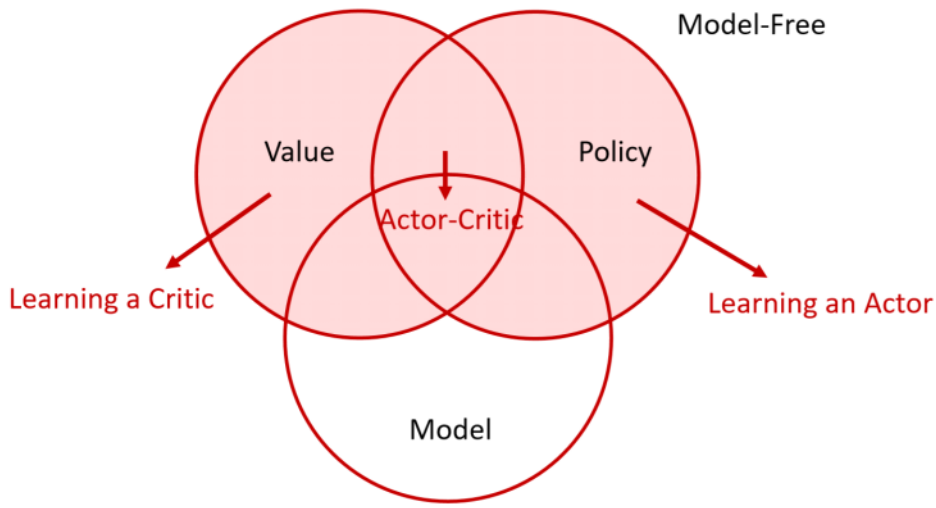
Model-based

- 對於環境去學
- 有model環境的變化

前面兩種已經隱含了環境的互動，所以表現通常比較好。學出action可以maxmize最終reward

explicit去學環境資訊不一定比較好

- Actor-Critic可以綜合前兩種優勢
- Model



Concluding Remarks

RL is a general purpose framework for **decision making** under interactions between *agent* and *environment*

- RL is for an *agent* with the capacity to *act* — action
- Each *action* influences the agent's future *state* — state
- Success is measured by a scalar *reward* signal — reward
- Goal: *select actions to maximize future reward*

An RL agent may include one or more of these components

- Value function: how good is each state and/or act
- Policy: agent's behavior function
- Model: agent's representation of the environmen