# 三维可视化助你直观理解DQN算法[DQN理论篇]

[](https://www.zhihu.com/people/zhang-si-jun-52)

[张斯俊](https://www.zhihu.com/people/zhang-si-jun-52" \t "/home/admin/文档\\x/_blank)

愿成为一把梯子，助你跨过无数的坑。

在上篇文章中，了解了深层神经网络到底是啥回事。现在把深层神经网络这套工具，用到强化学习中。

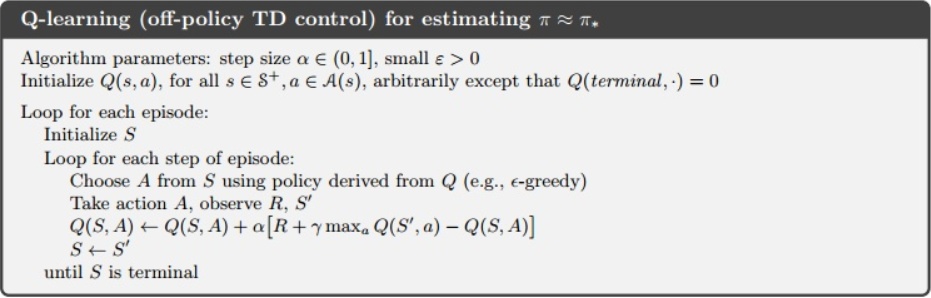
[张斯俊：一篇文章带你了解深度神经网络​](https://zhuanlan.zhihu.com/p/110531783" \t "/home/admin/文档\\x/_blank)

回想一下Qleaning，需要解决连续状态的问题。那么能否用神经网络解决呢？这就是DQN的思路。

如果还没了解Qlearning算法，建议先看一下之前的文章：

[张斯俊：[理论篇]怎样直观理解Qlearning算法？​](https://zhuanlan.zhihu.com/p/110338833" \t "/home/admin/文档\\x/_blank)

[张斯俊：手把手教你实现Qlearning算法[实战](https://zhuanlan.zhihu.com/p/110410276" \t "/home/admin/文档\\x/_blank)篇]

[](https://zhuanlan.zhihu.com/p/110410276)

### Deep network + Qlearning = DQN

先回顾一下Qlearning。

在Qlearning中，有一个Qtable，记录着在每一个状态下，各个动作的Q值。

Qtable的作用是当输入状态S，通过**查表**返回能够获得最大Q值的动作A。也就是需要找一个S-A的对应关系。

这种方式很适合格子游戏。因为格子游戏中的每一个格子就是一个状态，但在现实生活中，很多状态并不是**离散**而是**连续**的。例如在GYM中经典的CartPole游戏，杆子的角度是**连续**而不是**离散**的。在Atari游戏中，状态也是连续的。

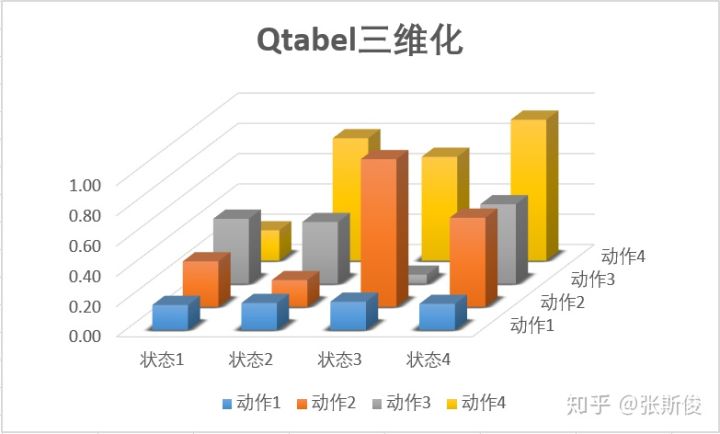
遇到这些情况，Qtable就没有办法解决。

刚才说了Qtable的作用就是找一个S-A的对应关系。所以就可以用一个函数F表示，有F(S) = A。这样就可以不用查表了，而且还有个好处，函数允许**连续**状态的表示。

这时候，深度神经网络就可以派上用场了。因为之前说过，神经网络就可以看成一个万能的函数。

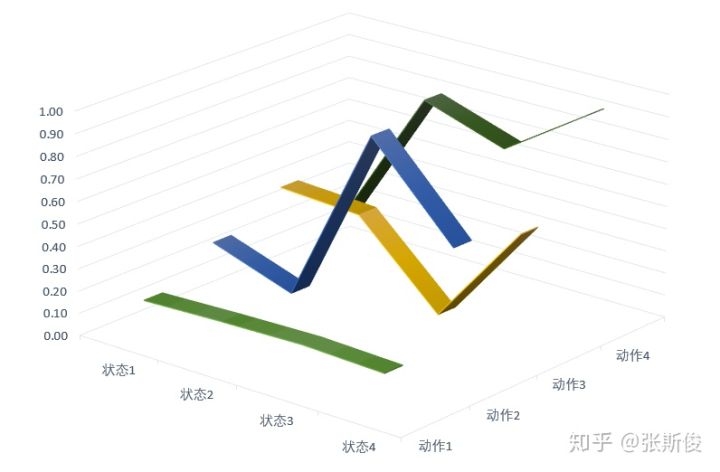
这个万能函数接受输入一个状态S，它能告诉我，每个动作的Q值是怎样的。

为了大家能更好理解神经网络的作用，可以从另外一个角度理解。 把Qtable三维可视化，就会得到这样一个图。

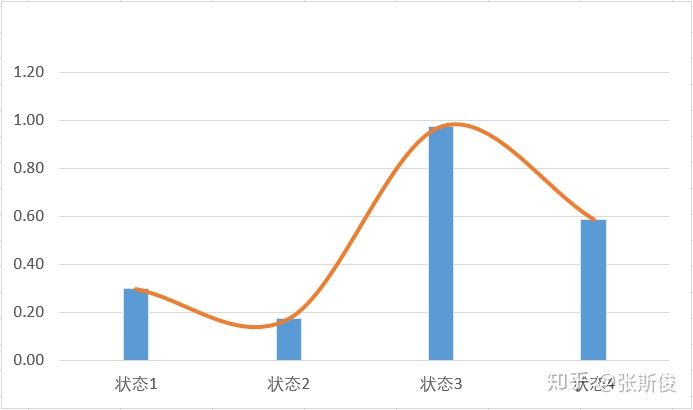


图中每根柱子的高度，表示状态S下，选择动作A的Q值。

现在用函数来表示，**相当于要扭曲一条曲线，这条曲线穿过了离散状态下的所有点。**



从二维的角度再看一下：



所以现在不但可以取状态3和状态4，还可以取状态3.5的Q值。

当然，在深度强化学习中，有很多数据的纬度相当高的。这里只是提供一个思考的角度。让大家能够可视化地理解DQN的思路。

其实DQN和Qlearning并没有根本的区别。只是DQN用神经网络，也就是一个函数替代了原来Qtable而已。

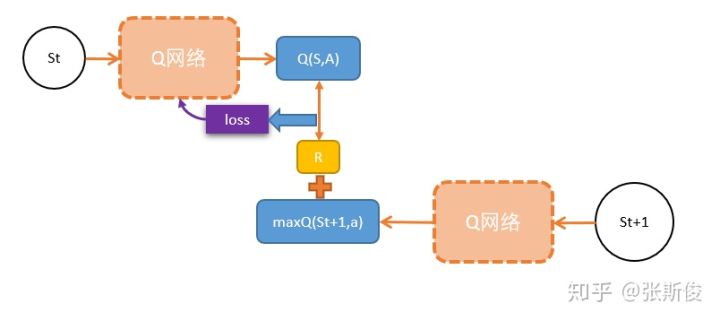
### 神经网络的目标

之前也说过，神经网络需要解决一个问题，就是更新目标怎么设置。

在手写数字识别等有监督学习的数据集中，有标签好的数据，也就是的目标是很明确的。那么在DQN中的目标是什么呢？

在Qlearning，用下一状态St+1的最大Q值替代St+1的V值。V(St+1)加上状态转移产生的奖励R就是Q(S,a)的更新目标。

来看下图：



假设需要更新当前状态St下的某动作A的Q值：Q(S,A),可以这样做：

1. 执行A，往前一步，到达St+1;
2. 把St+1输入Q网络，计算St+1下所有动作的Q值；
3. 获得最大的Q值加上奖励R作为更新目标；
4. 计算损失:

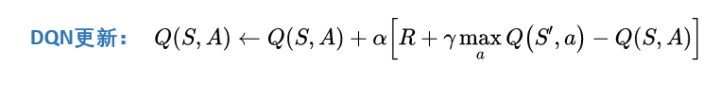
(1) Q(S,A)相当于有监督学习中的logits

(2) maxQ(St+1) + R 相当于有监督学习中的lables

(3) 用mse函数，得出两者的loss

1. 用loss更新Q网络。

也就是，用Q网络估算出来的两个相邻状态的Q值，他们之间的距离，就是一个r的距离。



### 总结

1. 其实DQN就是Qlearning算法中扔掉Qtable，换上深度神经网络。
2. 解决连续型问题，如果表格不能表示，就用函数，而最好的函数就是深度神经网络。
3. 和有监督学习不同，深度强化学习中，需要自己找更新目标。通常在马尔科夫链体系下，两个相邻状态状态差一个奖励r经常能被利用。

DQN其实没有什么神秘的，不是吗？

其实DQN在实现上，还有一些小技巧。将会在下一篇DQN实战篇看到。