**一、基本概念**

强化学习方法的基本做法是把样本输入LSTM神经网络，输出一个所有action的概率分布，把sample的action继续输入LSTM，直到完成一次决策，与ground\_truth的决策结果比较得到一个reward；上述是基本的policy网络（还可加入value网络作为baseline）

**二、与一般用海量样本的梯度下降法优化神经网络损失函数方法（记为W方法）的区别：**

强化学习最核心的作用是用来做决策，一个决策往往分很多步骤，比如一盘棋中有很多步落子。

每一步落子，在W方法中都对应一个有ground\_truth的样本，

而在强化学习解决的决策问题中，并不明确知道这每一步落子对应的ground\_truth,只知道一个决策做完后（一盘棋下完后）得到的reward；

因此，我们说**强化学习与W方法的区别在于样本少，需要在测试阶段采集**，意思是说强化学习的policy网络需要把每一步的输出作为下一步的输入完成一次完整的决策，称为测试过程，才能得到所谓的训练样本，进行梯度下降；

而W方法由于每一步的输出都有对应的ground\_turth,所以可以把前一步输出的ground\_truth作为下一步的输入而不是把前一步的输出作为下一步的输入，这样可以同时进行多步的计算，比较省时；因此，我们说强化学习比W方法训练更困难；

而具体应用在image caption问题中时，由于caption是一个句子，如果把句子中的每一个词作为每一步的ground-truth，则可以用W方法进行训练，如果把一个完整的句子作为得到reward的依据，则可以用强化学习的方法训练；

总而言之，并不是能用强化学习方法解决的问题都能用W方法解决，比如下棋，我们没法知道每一步落子对应的ground\_truth,也并不是能用W方法训练的问题都适合用强化学习的方法解决，强化学习更适用于解决由一个sequential组成一个完整policy的应用场合；