**标题**：Deep Reinforcement Learning with a Natural Language Action Space

**来源**：ACL 2016

**问题**：实验任务----文字游戏，实验目标----提出一种效果更好的DQN网络结构

1. 本文实验的目的是探索深度增强学习在nlp领域的应用，实验任务采用了nlp中较方便建立马尔科夫过程的文字游戏任务，文字游戏任务本质还是游戏，只是将当前游戏的环境和现状用文字描述出来，然后玩家根据阅读后的理解输入相应文字来作为操作方案。例如游戏显示“当前有两条岔路，请问选择左还是右？”，那么玩家则相应输入“左”或者“右”来作为操作。
2. 作者认为传统的DQN网络在解决上述问题时实验结果尚存不足，因而提出一个分开学习state和action向量表示的网络结构，并进行相应实验

**主要方法**：

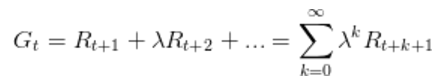
1. 本文采用的方法本质是深度增强学习，下面简单介绍一下：

深度增强学习即增强学习在深度学习上基础的实现，本质是用深度神经网络来拟合增强学习中的Q函数，带来的优势是更强的拟合效果以及引入端到端的解决能力。增强学习的建模过程是基于马尔可夫过程建模，因而训练目标有别于传统机器学习任务，传统任务的训练目标是减小期望输出和目标输出的绝对误差，拟合过程是在单个时间点，而增强学习是寻求一个长期增益最大，且长期奖励结果只存在相对最优，拟合过程也是建立在一段时间上。

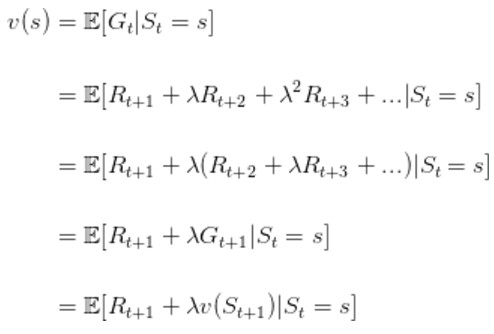
增强学习有三个重要参数，状态state，动作action，奖励reward。整个任务的行进是基于马尔可夫过程的时间序列，即某个state下，采取某个action，获得相应的reward。



优化目标是使采取当前动作后，未来的累积奖励最大，其中λ是折扣因子



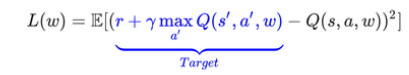
将上式经过等价变换后可以得到一个服从迭代过程的bellman方程



等式中状态价值函数v(s)可等价替换为动作价值函数Q，当Q表示的是最优策略的价值时即为如下等式



因而优化目标即让当前策略的动作价值函数无线逼近最优策略的价值函数，所以loss函数如下



至此，整个深度增强学习网络的优化目标就出来了，在这里讲解的比较简略，后文的理解大家只需要套用深度学习的训练，然后误差函数如上设置即可，其中的w即为网络权重。如果还有什么疑问欢迎自己查阅资料。

（2）本文具体工作

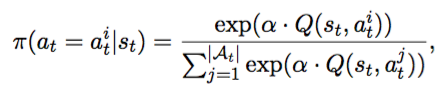
(a)运用mlp分别构建两个network,并各自将文本描述的状态和动作encode为一个向量然后用一种逐元素操作方法得到Q值，比如对两个向量作内积。内积就是相似度的一种表征，也就是本文模型中的relevant。

(b)然后运行值更新策略，更新策略网络Q，这里的Q网络即上一步由两个mlp merge出来的模型结构。



当DQN网络策略最优时，上式趋于0。本文并未提到奖励如何得来，也许由游戏本身反馈。

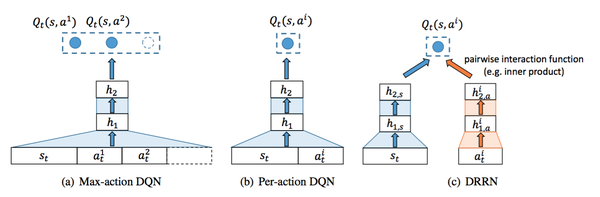
(c)探索与利用机制，本文并未采用传统的e-greedy，而是采用softmax selection



上式的α称为温度，当它趋于0时倾向“仅利用”策略，当它趋于无穷时倾向“仅探索”，在训练过程中该参数会逐渐由大变小。

**相关工作：**

1. 举出另外两种baseline模型与本文模型比较



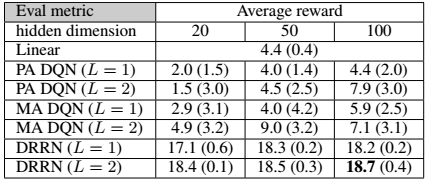
(a) Max-action DQN

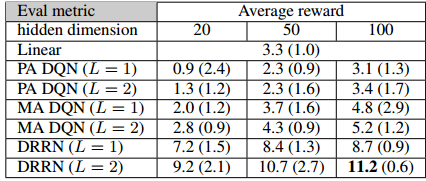
该模型适用于每一个transition中actions的最大数量是已知的情况，将每个transition中state和actions拼接成一个向量作为输入，通过一个Deep Network得到每个action的Q值。论文Generating Text with Deep Reinforcement Learning采用的就是这种网络结构

(b) Per-action DQN

该模型将每一对(state,action)拼接成一个向量作为输入，通过network得到每个action的Q值。

（2）在Saving John和Machine of Death两个文字游戏上的实验结果





**简评：**

1. 本文的核心是将深度增强学习应用到广义上的nlp领域，但它的亮点则在于提出的DRRN模型。（我觉得text-game严格意义上来说不能算nlp任务）

首先需要指出率先将DQN应用到广义nlp领域的文章是Language Understanding for Text-based Games using Deep Reinforcement Learning，采用的模型是简化版的Max-action DQN，即没有输入候选action。率先将DQN应用到严格意义上nlp领域的文章是Generating Text with Deep Reinforcement Learning。两者的模型都是MA DQN，action也都是词，action-size也就是词表的大小，那么必然面临一个nlp任务中常面临的大词表问题。词表越大，难收敛问题也就越严重。

而本文的DRRN模型相比其他模型的优势就在于单独用两个网络分别映射state和action.不仅解决了大词表问题，而且降低了网络规模，并增加了网络信息容量。而且因为state往往是一个比较长的文本，可能是几句话，而action一般来说是一个动词短语，如果将state和action拼接送入单个网络结构后会降低action的影响，所以将state和action通过不同的网络结构进行学习，得到相同维度的表示再结合计算Q值可能效果更好。这也在实验结果中得到了体现。

然而DRRN模型虽然名字很新，但它其实依旧是一个传统的retrieval模型，只不过在DQN领域摇身一变穿了新衣，这种retrieval模型在QA领域极其常见，即将问题和答案分别用RNN encode出一个向量然后计算相似度，然后根据相似度打分rank出最优结果。

（2）增强学习和nlp

在迭代过程中有两个重要参数是r和Q，前者是奖励，后者是动作价值函数，它的值反映的是长期的奖励累积。本文并没有指出r的由来，应该是由游戏直接反馈结果，然而实际任务可能不存在这个反馈，所以就需要我们自己去生成对应的r值。例如，著名的alphaGo里有两个主要的网络结构，值网络和策略网络，前者就是对应的r。在nlp中，例如seq2seq任务，r值就可以是BLEU，QA中，r值就可以是NDCG。增强学习的优化目标是Q，既追求的是一个长期过程，建模过程也用了马尔科夫决策建模，并不像传统的机器学习任务那样，每个样本是一个独立的完整事件。所以当我们在将增强学习运用到nlp领域时，我们需要思考什么样的nlp任务是这种满足长期累积的平稳过程，或者怎么将nlp任务转化为这种长期任务来处理。

1. 延伸阅读

Narasimhan K, Kulkarni T, Barzilay R. Language Understanding for Text-based Games Using Deep Reinforcement Learning[J]. EMNLP, 2015, 40(4):1-5.

Guo H. Generating Text with Deep Reinforcement Learning[J]. NIPS, 2015.

Bahdanau D, Brakel P, Xu K, et al. An Actor-Critic Algorithm for Sequence Prediction[J]. arXiv preprint arXiv:1607.07086, 2016.