受行为驱动学科的启发，强化学习是关于智能体与环境交互的一种学习方式。环境处于一定的状态下，智能体通过一种行为与环境产生互动，然后环境给与智能体奖励，如何最大化这种奖励就是强化学习要解决的问题。20世纪末，伴随着机器人应用的研究发展，强化学习也逐渐称为热门的研究领域。特别是近十几年来随着深度学习的研究，越来越多的研究与深度学习相结合取得了突破性的进展，如Hinton等用RBM[1]深度神经网络实现对图像的降维，取得比PCA[2]更好的效果，被誉为首次深度学习兴起的成功实践。强化学习也不例外，通过与深度学习的结合，解决了很多复杂的连续状态空间、连续动作空间的问题，为机器人、金融、生物工程等领域开创了新局面。

强化学习可以分为基于模型的学习（model-base）和无模型的学习（model-free）两大类。基于模型的算法需要预先知道环境的状态转移函数和奖励函数，或者可以根据智能体与环境的交互采样数据学习到，譬如用动态规划的策略迭代和价值迭代算法，经典的Dyna-Q[3]算法等。无模型的算法则相反，不知道环境的模型参数，而是直接通过智能体与环境的交互数据直接学习策略或状态价值，如DQN[4]、策略梯度[5]、DDPG[6]、PPO[7]、SAC[8]等，都是这类型的算法。两种学习方法各有优缺点。通常在确定性环境中会使用基于模型的学习算法，如某些具有严格规则的棋牌类游戏。但这种白盒环境在现实中很少，很难对这种复杂环境建立良好的模型，这时使用无模型的学习算法会更加容易训练。 本文介绍的基于值函数的算法也都是属于无模型类型的学习算法。

最初关于值函数的表格型方法，譬如Q-learning^{[9]}[9]，Sarsa^[10]{[10]}，只是解决离散状态下的离散动作问题。随着状态空间的增大，需要引入深度神经网络来“代替”这个表格。所以DQN全称为Deep Q Network^{[4]}，可以用来解决连续状态下的离散动作问题。为了逼近真实的值函数，可以将平方误差作为DQN的目标函数：

通过实践可以看到光引入深度神经网络，网络的训练会非常不稳定。因为网络的参数不断迭代变化，计算目标的Q值也是不断变化，试想想，要逼近一个随时在变化的目标是十分困难的。 于是，Volodymyr Mnih[11]等^{[11]}提出了使用两个相同结构的神经网络f\_\theta和f\_{\theta^-}架构来组织DQN：

Tom Schau[12]l等^{[12]}发现认为TD误差大（TD\_{error} = Q\_\theta(s\_i, a\_i) - (r\_i+ \gamma \max\limits\_{a^{'} \in A}Q\_\theta(s\_i^{'}, a\_i^{'}))）的样本应该给予更高的权重，提出了基于优先级的经验回放方法，使用TD误差作为样本采样的优先级，用非均匀采样替代了之前的均匀采样，获得了不错的训练效果。

普通的DQN中选择动作和计算动作Q值都是用同一个网络，很容易造成Q值的高估。实际上因为强化学习中总是进行最大化价值的操作，Q值高估的现象十分常见。Hado van Hasselt等提出的Double DQN[13]^{[13]}（简称DDQN）致力于解决这一问题。DDQN选择动作用的时训练用的网络f\_\theta，而计算该动作价值的则用目标网络f\_{\theta^-}，由于目标网络f\_{\theta^-}的参数更新不会很频繁，一定程度抑制的Q值的过大。图2是论文给出的分别使用DQN和DDQN进行57个Atari游戏的得分情况，实践证明DDQN比普通的DQN性能有所提升。

Wangzi Yu等提出的Dueling DQN[14]^{[14]}也是在DQN的基础上进行改进，以改善Q值高估的问题。论文指出，根据优势函数表示采取不同动作差异性的意义：A(s, a) = Q(s, a) + V(s)，神经网络不直接输出Q值，Dueling DQN改为输出状态价值（V值）和优势值（A值），再求和得出Q值，网络架构改动如图3所示。

于是提出了NoiseNet[15]^{[15]}方法，在神经网络的权重上添加参数化的噪声，以此提高算法探索的效率。论文指出，在神经网络权重加入噪声带来的不确定性，比在策略上的更加大，而且噪声的参数也是需要学习的，网络可以通过学习来调整噪声的大小。

LSTM+DQN的DRQN[16]^{[16]}方法用于解决这种部分观测的马尔科夫决策过程问题。作者将DQN网络最后一层全连接层改为LSTM，最终网络结构如图4。

Ivan Sorokin等在此基础上增加注意力机制，提出了DARQN[17]^{[17]}，使得LSTM层不仅要选择下一个动作，还要选择下一个动作要关注的区域，进一步增强了时序强化学习的可解释性。

Deepmind和MIT提出一种分层强化学习方法H-DQN[18]

Rainbow[19]^{[19]}是Deepmind提出的集DQN之大成的方法。它主要使用了如下技巧：

深度强化学习在各人领域都取得重大成功。从早期的Atari系列游戏智能操控，到近几年来的AlphaGo利用深度神经网络和蒙特卡罗树搜索击败国际象棋冠军，还有前几年Atlas波士顿动力机器人的仿人类行为，无不体现人们对强化学习越来越感兴趣。

机器人技术一直是强化学习的深耕领域。深度强化学习可以在现实环境中实现复杂的机器人行为控制，譬如堆叠积木、拧瓶盖、跑酷等。譬如Andy Zeng等[20]利用深度强化学习训练抓娃娃机，通过摄像头捕获实物图像，输入到DQN进行训练，网络输出一张关于Q值分布的热力图，图中数值最大的点就是机械臂垂直向下移动的位置。Al-Nima等[21]通过手机汽车行驶时的视图作为输入状态，基于深度强化学习生成了适合的汽车道路跟踪行为。还有Nagabandi等[22]提出一种学习适应动态现实环境的方法，进一步缓解了训练数据缺失的问题，并在处理机器人操作任务是具有更好的泛化能力。

在自然语言处理方面，深度强化学习也有着广泛的应用前景，并已经成功应用于神经机器翻译、对话系统、语言生成等领域。Jiaotao Gu等[23]采用迁移学习和深层元强化学习的方法，将多种源语言中的词汇和句子表达迁移到一种目标语言中。他们还首次将深度元强化学习算法扩展到低资源的神经机器翻译[24]，结果表明，该方法的性能明显优于以往的多语言方法，只需少量的训练样本就可以训练处具有竞争力的神经机器翻译系统。Yutian Chen等[25]提出一种基于深度元强化学习的自适应文字合成语音方法，该方法使用很少的语言样本在新场景中可以高还原度回复说话人的声音。

当然，深度强化学习目前还存在着相当多的挑战。譬如：

* 需要从有限的样本中学习真实环境。这也就是样本利用率比较低效的问题。要训练一个误差较低的DQN时，目前确实需要十分大量的样本，而且需要相当多的回合数，譬如Rainbow【hoho:】就需要大概83小时的训练才能超过人类水平。目前研究会逐渐更多的考虑使用model-base和model-free向结合的方法，应用model-base的方法已经知道的环境的模型参数了，不需要大量的训练样本也能有比较好的效果。
* 奖励函数未知或者具有多重目标的奖励。人们并不确切知道自己想要优化哪方面的收益，所以此时奖励函数往往是多维的，仅仅用各个维度的均值作为收益是不准确的。然后定义一个奖励函数通常十分困难，这需要足够的先验知识，需要对所有的状态有所把握。目前可以借助使用逆强化学习或模仿学习，它们都不需要奖励函数。
* 信用分配问题。如何分配系多智能体对结果的贡献，分析动作序列对最终奖励的贡献，有助于加快学习速度，同时也有助于网络的可解析性发展。

**参考文献**

[1] G. E. Hinton,R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks[Z], 2006.

[2] Karl Pearson. On lines and planes of closest fit to systems of points in space[Z], 1901.

[3] Richard S.Sutton. Dyna, an Integrated Architecture for Learning, Planning, and Reacting[C], 1991.

[4] Volodymyr Mnih,Koray Kavukcuoglu,David Silver, et al. Playing Atari with Deep Reinforcement Learning[C]//Nips, 2013.

[5] Richard S. Sutton,David McAllester,Satinder Singh, et al. Policy Gradient Methods for Reinforcement Learning with Function Approximation[Z], 1999.

[6] David Silver,Guy Lever,Nicolas Heess, et al. Deterministic Policy Gradient Algorithms[C]//Pmlr, 2014.

[7] John Schulman,Filip Wolski,Prafulla Dhariwal, et al. Proximal Policy Optimization Algorithms[Z], 2017.

[8] Tuomas Haarnoja,Aurick Zhou,Kristian Hartikainen, et al. Soft Actor-Critic Algorithms and Applications[Z], 2018.

[9] CHRISTOPHER J.C.H. WATKINS. Q-Learning[Z], 1992.

[10] G. A. Rummery,M. Niranjan. Generalization in reinforcement learning: Successful examples using sparse coarse coding[Z], 1994.

[11] Volodymyr Mnih,Koray Kavukcuoglu,David Silver, et al. Human-level control through deep reinforcement learning[J], 2015.

[12] Tom Schaul,John Quan,Ioannis Antonoglou, et al. Prioritized Experience Replay[C]//Iclr, 2016.

[13] Hado van Hasselt,Arthur Guez,David Silver. Deep Reinforcement Learning with Double Q-learning[C]//Aaai, 2016.

[14] Ziyu Wang,Tom Schaul,Matteo Hessel, et al. Dueling Network Architectures for Deep Reinforcement Learning[C]//Pmlr, 2016.

[15] Meire Fortunato,Mohammad Gheshlaghi Azar,Bilal Piot, et al. Noisy networks for exploration[C]//Iclr, 2018.

[16] Matthew Hausknecht,Peter Stone. Deep Recurrent Q-Learning for Partially Observable MDPs[C]//Aaai, 2015.

[17] Ivan Sorokin,Alexey Seleznev,Mikhail Pavlov, et al. Deep Attention Recurrent Q-Network[C]//Nips, 2015.

[18] Tejas D. Kulkarni,Karthik Narasimhan,Ardavan Saeedi, et al. Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation[C]//Nips, 2016.

[19] M Hessel,J Modayil,H Van Hasselt. Rainbow: Combining Improvements in Deep Reinforcement Learning[C]//Aaai, 2018.

[20] Andy Zeng,Shuran Song,Stefan Welker, et al. Learning Synergies between Pushing and Grasping with Self-supervised Deep Reinforcement Learning[R], 2018.

[21] Raid Rafi Omar Al-Nima,Tingting Han,Taolue Chen. Road Tracking Using Deep Reinforcement Learning for Self-driving Car Applications[Z], 2019.

[22] Anusha Nagabandi,Ignasi Clavera,Simin Liu, et al. :Learning to adapt in dynamic, real-world environments through meta-reinforcement learning[Z], 2018.

[23] Jiatao Gu,Hany Hassan,Jacob Devlin, et al. Universal Neural Machine Translation for Extremely Low Resource Languages[C]//Naacl, 2018.

[24] Jiatao Gu,Yong Wang,Yun Chen, et al. Meta-learning for low-resource neural machine translation[C]//Emnlp, 2018.

[25] Yutian Chen,Yannis Assael,Brendan Shillingford, et al. Sample eﬃcient adaptive text-to-speech[C]//Iclr, 2018.