Playing Atari with Deep Reinforcement Learning

&

Human-level control through deep reinforcement learning

技术报告

人工智能 91 卢佳源 2191121196

一、 论文试图解决什么问题?

本文试图应用强化学习方法让深度学习模型直接从高位感官输入学习控制策略;（目标是将一个强化学习算法连接到一个深度神经网络，该网络直接运行在 RGB 图像上，并通过使用随机梯度更新有效地处理训练数据。）

a) 高维特征 VS 基于手工提取特征且线性表示的传统 RL 方法:
   i. 基于手工提取特征且线性表示的传统 RL 方法:
      1. 截至当时，“直接从视觉和语音等高维感官输入中学习如何控制一个 Agent”这个问题对强化学习来说是一个长期面临的挑战，因为当前的 RL 应用大多依赖手工提取的特征以及线性价值函数或者策略表示,因此其性能很大程度上依赖特征表示的质量;另外，现实中的非线性关系很难用线性方程表示;
   ii. 高维特征:
      1. 当前深度学习已经有了一定的发展,可以运用神经网络解决高维特征的问题，因此很自然联想到将深度学习运用于 RL;
b) 单纯的深度学习思想 VS 传统 RL 思想:
   i. 训练数据的标定:
      1. 深度学习是有监督学习，需要大量的手工标定的训练数据，且输入和目标之间是直接关联的;
      2. 而 RL 方法需要从一个通常是稀疏、有噪声和演示的标量奖励信号中学习，即行动和结果的奖励之间可能有很长的延时;
   ii. 数据样本的独立性:
      1. 深度学习要求样本是独立（同分布）的，因此深度学习可以基于一个固定的底层分布来设计模型;
      2. RL 中却经常会遇到高度相关的状态序列，且数据分布会随着学习的进行而发生改变，无法直接套用深度学习模型;

二、 这是否是一个新的问题?

a) 这是一个新的问题，2013 年的文章是第一篇将深度学习应用于强化学习方法的研究，2015 年的文章是在此基础上的改进。

三、 这篇文章要验证一个什么科学假设？

a) 本文的目标是创建一个单一的神经网络 Agent，使其可以成功地学习去玩更多种类的游戏；
b) 本文使用的网络完全模拟人类玩家，即不具有任何特定的游戏的信息或者手工设计的视觉功能，以及对模拟器内部的状态也是未知的，它仅仅从视频的输入、奖励和终端信号以及可能的动作中学习；
c) 评估网络架构模拟的学习与人类学习的性能，即验证 DRL 在游戏中得到的成绩是否比人类得到的成绩有所提升，或基本上能够模拟人类的学习过程。

四、 有哪些相关研究？如何归类？谁是这一课题在领域内值得关注的研究员？

a) TD-gammon（TD-双陆棋）：
[24] Gerald Tesauro. Temporal difference learning and td-gammon. Communications of the ACM,38(3):58–68, 1995.
   i. 是一个完全通过强化学习和自我游戏来学习的反向双陆棋程序，且性能超过人类玩家；
   ii. 使用了一种类似 Q-Learning 的无模型强化学习算法，并使用带有一个隐藏层的 MLP 来逼近价值函数；
   iii. 缺点：只适用于西洋双陆棋，而无法应用于国际象棋、围棋和跳棋，可能的原因是西洋双陆棋所用的骰子滚动的随机性有助于状态空间的探索，使得价值函数的逼近较为平滑。

b) 无模型强化学习算法和非线性函数逼近器（或者与非策略学习）相结合的方法：
[25] John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. Automatic Control, IEEE Transactions on, 42(5):674–690, 1997.
   i. 可能会导致 Q-network 发散；
   ii. 研究具有更好收敛性的线性函数逼近器，使得 Q-network 的收敛性得以保证；

c) 深度学习与强化学习相结合：
[21] Brian Sallans and Geoffrey E. Hinton. Reinforcement learning with factored states and actions. Journal of Machine Learning Research, 5:1063–1088, 2004.
   i. 深度神经网络：估计环境ε；
   ii. 限制的玻尔兹曼机：估计价值函数（或者策略）；
   iii. 梯度时间差分方法 TD：部分解决了 Q-Learning 的差异问题；
   iv. 可以在非线性函数逼近器评估固定策略时收敛；
   v. 缺点：还没有推广到非线性控制。

d) 神经拟合的 Q-Learning（NFQ）：
[12] Sascha Lange and Martin Riedmiller. Deep auto-encoder neural networks in reinforcement learning. In Neural Networks (IJCNN), The 2010 International Joint Conference on, pages 1–8. IEEE, 2010.

i. NFQ 优化了损失函数序列，使用 RPROP 算法来更新 Q-network 的参数，每次迭代的恒定成本较批处理更新要低；

ii. 成功地应用于纯视觉输入的简单现实世界控制任务：
   1. 首先使用深度自动编码器学习任务的低维表示；
   2. 然后将 NFQ 应用于上诉的低维表示；

iii. 特点：
   1. 直接从视觉输入进行端到端的应用强化学习；
   2. 因此可以学习与区分动作价值直接相关的特征；

e) 具有线性函数逼近和通用视觉特征的标准强化学习算法：
[3] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. Journal of Artificial Intelligence Research, 47:253–279, 2013.

i. 使用更多的特征，和 tug-of-war 哈希将特征随机投影到一个低维空间中，使结果得到改进；

f) HyperNEAT 进化结构：用于进化一个代表该游戏策略的神经网络：
[8] Matthew Hausknecht, Risto Miikkulainen, and Peter Stone. A neuro-evolution approach to general atari game playing. 2013.

i. 当使用模拟器的重置工具对确定性序列进行反复训练时，这些策略能够利用几个 Atari 游戏中的设计缺陷。

五、 论文中提到的解决方案之关键是什么？

a) DQN=CNN+Q-Learning：

i. Model-free+off-policy；

ii. 输入为原始图像；输出为每个可能的动作对应的动作价值函数来估计未来的回报（用 CNN 逼近最优动作价值函数）；

iii. 具体算法流程：

Algorithm 1: deep Q-learning with experience replay.

Initialize replay memory D to capacity N

Initialize action-value function Q with random weights θ

Initialize target action-value function $\hat{Q}$ with weights $\theta^{-} = \theta$

For episode = 1, M **do**

  Initialize sequence $s_1 = \{x_1\}$ and preprocessed sequence $\phi_1 = \phi(s_1)$

  For t = 1,T **do**

    With probability ε select a random action a_t

    otherwise select $a_t = \text{argmax}_a Q(\phi(s_t), a; \theta)$

    Execute action a_t in emulator and observe reward r_t and image x_{t+1}

    Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$

    Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in D

    Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from D

    Set $y_j = \begin{cases} r_j & \text{if episode terminates at step j+1} \\ r_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^{-}) & \text{otherwise} \end{cases}$

    Perform a gradient descent step on $\left(y_j - Q(\phi_j, a_j; \theta)\right)^2$ with respect to the network parameters θ

    Every C steps reset $\hat{Q} = Q$

  End For

End For

b) 经验回放：
   i. 目的：解决数据样本序列的强相关性导致 Q 的微笑改变就会导致策略的很大的变化的问题；
   ii. 使用了一种称为经验回放地级数存储每一个时刻的 Agent 的经验，将 Q-Learning 更新或者小批量更新应用于从存储样本的池子中随机提取的经验样本；
   iii. 在执行完经验回放后，Agent 更加ε-greedy 算法选择并执行一个操作。
   iv. 经验回放的优点：
      1. 经验的每一步都有可能用于许多权重的更新，使得被允许的数据效率更大；
      2. 直接从连续样本进行的学习是低效的，因为连续样本之间的相关性很强，而 DQN 使用的样本随机后打破了样本之间的相关性，从而减少了更新的方差；
      3. 在学习 on-policy 时，当前参数决定了参数被训练的下一个数据样本，但容易导致陷入局部最小值；
      4. 使用了经验回放，使得动作分布在其之前的许多状态上被平均，平滑了学习，避免了参数中的震荡或者发散。
c) 帧跳过技术：Agent 在每 k 帧选择动作，而非每一帧上选择动作，其最后一个动作在跳过的帧上重复；
d) 使用了另外一个单独的网络来生成目标，使其适用于不会发散的大型神经网络的训练，解决了 2013 年论文出现的长时间尺度问题。

六、 论文中的实验是如何设计的？

a) 预处理和模型架构：
   i. 预处理：
      1. 目的：降低输入维数；
      2. 步骤：
         a) 将 RGB 表示的图像转换为灰度图像；
         b) 将原始的 210*160 像素图像降采样到 110*84 像素；
         c) 通过裁剪图像的 84*84 的区域来获得最终输入模型的表示；
   ii. 模型架构：
      1. 对每个可能的动作都有一个单独的输出单元，并且只有状态表示是神经网络的输入；
      2. 输出对应于输入状态的单个动作的预测的 Q 值；
      3. 优点：能够计算在给定状态下的所有可能的动作的 Q 值；
      4. 神经网络 DQN 的每一层（2013 年）：
         a) 输入：84*84*4 的图像；
         b) 第一个隐藏层：16 个步幅为 4 的 8*8 滤波器与输入图像进行卷积，并使用整流非线性单元进行激活；
         c) 第二个隐藏层：32 个 4*4 的滤波器，用整流非线性单元激活；
         d) 第三个隐藏层：全连接层，由 256 个整流单元组成；
         e) 输出层：全连接线性层，每个有效的动作都有单个输出。
      5. 神经网络 DQN 的每一层（2015 年）：

a) 输入：84*84*4 的图像；
b) 第一个隐藏层：32 个步长为 4 的 8*8 滤波器与输入图像卷积，并使用整流非线性单元激活；
c) 第二个隐藏层：64 个步长为 2 的 4*4 滤波器与该层的输入卷积，并使用整流非线性单元激活；
d) 第三个卷积层：64 个步长为 1 的 3*3 滤波器，并使用 512 个整流单元激活；
e) 输出层：是一个全连接线性层，使得每个有效的动作都有单独的输出；

b) 实验过程：
i. 使用批大小为 32 的小批量 RMSProp 算法；
ii. 训练过程中的动作策略选择使用ε-greedy 算法；
iii. 总共训练 1000 万帧，并用其最近的 100 万帧进行经验回放；
iv. 帧跳过技术：Agent 在每 k 帧选择动作，而非每一帧上选择动作，其最后一个动作在跳过的帧上重复。

七、 用于定量评估的数据集是什么？代码有没有开源？

a) 用于定量评估的数据集：
i. 7 款流行的 ATARI 游戏（包括 Beam Rider, Breakout, Enduro, Pong, Q*bert, Seaquest, Space Invaders）；
ii. 在 7 款游戏中，使用相同的网络架构、学习算法和超参数设置，在不包含特定游戏信息的情况下处理各种游戏；
iii. 仅仅在训练中对游戏的奖励结构进行一个改变：
1. 由于不同游戏的分数规模差异较大，因此将所有的正奖励固定为 1、负奖励固定为-1，奖励不变为 0；
2. 优点：限制了误差导数的规模，并更容易在多个游戏中使用相同的学习率；

b) 2013 年的文章代码没有开源，2015 年的文章代码开源了：
https://sites.google.com/a/deepmind.com/dqn for non-commercial uses only.

八、 论文中的实验及结果有没有很好地支持需要验证的科学假设？

a) 2013 年论文的实验结果：
i. 图 2：在训练过程中预测的 Q 值相对平稳地提升（避免了小的 Q 值变化使策略发生较大改变的问题），表明本文地方法能够使用强化学习信号和随机梯度下降以稳定地方式训练大型神经网络；
ii. 图 3：游戏网格上学习地价值函数地可视化：图中显示，在敌人出现在屏幕地左侧点 A 时，价值 Q 的预测值会跳跃；当特工向敌人发射一枚鱼雷并即将几种敌人时（点 B），Q 的预测值达到最大峰值；在敌人消失后（点 C），Q 的预测值大致下降到原始值。表明本文的方法能够学习价值函数在一个合理复杂的事件序列中如何演变的。
iii. 本文的方法在 Breakout, Enduro, Pong 这三个游戏上取得了比人类专家玩家更好的性能，在 Beam Rider 上取得了接近人类玩家的性能，但在 Q*bert, Seaquest,

Space Invaders 上还距离人类玩家能够达到的性能很远，因为这三个游戏需要网络找到一个扩展长时间尺度的策略，更具有挑战性。

b) 2013 年的论文很好地支持需要验证地科学假设："将一个强化学习算法连接到一个深度神经网络，该网络直接运行在 RGB 图像上，并通过使用随机梯度更新有效地处理训练数据"。但对于能够找到一个扩展长时间尺度的策略还是仍有不足。

c) 2015 年论文的实验结果：DQN Agent 在 49 款游戏中的表现与专业人类玩家的水平相当，在超过一半的游戏中获得了超过 75%的人类分数(即 29 场游戏)；

d) 2015 年论文在 2013 年论文的基础上解决了长时间尺度的问题，使得 Agent 性能有了更大的提升。

九、 这篇论文到底有什么贡献？

a) 2013 年的论文提出了一种新的强化学习深度学习模型，并展示了在仅使用原始像素作为输入时，能够掌握 Atari2600 游戏的困难控制策略的能力；

b) 2013 年的论文还提出了一种在线 Q-Learning 的变体，即结合律随机小批量更新和经验回放，来简化对 RL 的深度网络的训练，即将机器学习技术和生物启发的机制相结合来让 Agent 能够处理更复杂的任务；

c) 2015 年的论文使用了另外一个单独的网络来生成目标，使算法更适用于大型长延时的神经网络的训练。

十、 下一步呢？有什么工作可以继续深入？

a) 在 2013 年的实验中，7 款中的 6 款游戏给出了最先进的结果，但仍有三个游戏没有超越人类玩家的水平，因此针对这三个游戏体现出来的需要长时间尺度的策略特征，之后可以进行改进和突破。

b) 在 2015 年的实验中，DQN 在 Atari2600 游戏中的 49 个游戏都超过了人类玩家（即使用了另外一个单独的网络来生成目标，使其适用于不会发散的大型神经网络的训练），未来的探索使经验回放内容偏向显著事件的潜在用途（模拟生物海马体中经验回放机制）。

c) 提升 CNN 网络和 Q-Learning 的收敛性；

d) 增强 DQN 中 Q-network 的记忆能力。