通过随机网络蒸馏法进行探索

尤里-布尔达
兴业银行

哈里森-爱德华兹
兴业银行

Amos Storkey
爱丁堡大学

奥列格-克里莫夫
敞开式人工智能

摘要

我们为深度强化学习方法引入了一个探索奖金，它易于实现，并为所进行的计算增加最小的开销。该奖励是神经网络预测由固定的随机初始化神经网络给出的观察结果的特征的误差。我们还介绍了一种灵活地结合内在和外在奖励的方法。我们发现，随机网络提炼（RND）奖励与这种增加的灵活性相结合，能够在几个难于探索的Atari游戏上取得重大进展。特别是我们在Montezuma's Revenge上建立了最先进的性能，这个游戏对深度强化学习方法来说是很难的。据我们所知，这是第一个在不使用演示或接触到游戏底层状态的情况下，在这个游戏上取得优于人类平均表现的方法，并偶尔完成了第一关。

1 简介

强化学习（RL）方法通过最大化策略的预期收益来工作。当环境有密集的奖励，通过采取随机的行动序列很容易找到时，这种方法很有效，但当奖励稀疏且难以找到时，往往会失败。在现实中，为希望RL代理解决的每一项任务设计密集的奖励函数往往是不切实际的。在这些情况下，有必要采用定向方式探索环境的方法。

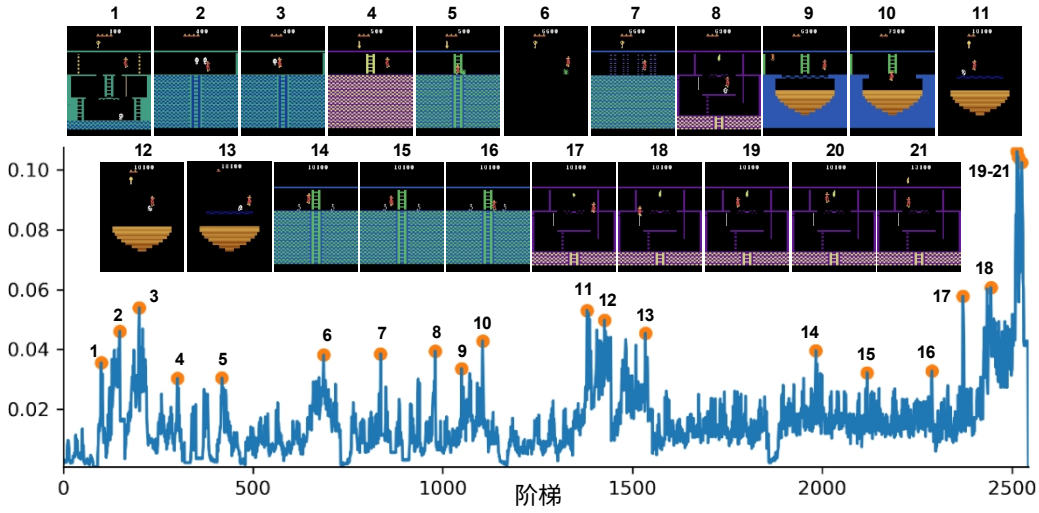


图1：RND探索奖金在第一集的过程中，特工拿起了火把（19-21）。为此，特工经过17个房间，收集宝石、钥匙、一把剑、一个护身符，并打开两扇门。探索奖励中的许多尖峰对应着有意义的事件：失去一条生命（2,8,10,21），险些被敌人逃脱（3,5,6,11,12,13,14,15），通过一个困难的障碍物（7,9,18），或捡到一个物体（20,21）。最后的大尖峰对应的是与火炬互动的新体验，而较小的尖峰对应的是相对罕见的事件，但代理人已经经历过多次。请看这里的视频。

*按字母顺序排列；前两位作者贡献相同。

RL的最新发展似乎表明，解决最具挑战性的任务（Silver等人，2016；Zoph & Le，2016；Horgan等人，2018；Espeholt等人，2018；OpenAI，2018；OpenAI等人，2018）需要处理从并行运行许多环境副本中获得的大量样本。有鉴于此，最好能有能在大量经验的情况下良好扩展的探索方法。然而，最近推出的许多基于计数、伪计数、信息增益或预测增益的探索方法很难扩展到大量的并行环境。

本文介绍了一种探索奖金，这种奖金实现起来特别简单，在高维观察中效果很好，可以与任何策略优化算法一起使用，而且计算效率高，因为它只需要神经网络在一批经验上进行一次前向传递。我们的探索奖金是基于这样的观察，即神经网络在类似于它们被训练过的例子上的预测误差往往明显较低。这促使我们使用在代理人过去经验上训练的网络的预测误差来量化新经验的新颖性。

正如许多作者所指出的，使这种预测误差最大化的代理人往往被吸引到预测问题的答案是输入的随机函数的过渡中。例如，如果预测问题是在当前观察和代理的行动（前向动力学）的情况下预测下一个观察，那么试图使这种预测误差最大化的代理将倾向于寻求随机的过渡，比如那些涉及随机变化的电视上的静态噪声，或者随机事件的结果，如抛硬币。这一观察促使人们使用量化预测的相对改进，而不是其绝对误差的方法。不幸的是，如前所述，这种方法很难有效实施。

我们为这种不理想的随机性提出了另一种解决方案，即使用预测问题来定义探索奖金，其中答案是其输入的确定性函数。也就是说，我们预测一个固定的随机初始化的神经网络在当前观察上的输出。

自Mnih等人（2013）的开创性工作以来，Atari游戏一直是深度强化学习算法的一个标准基准。Bellemare等人（2016）在这些游戏中确定了具有稀疏奖励的困难探索游戏：Freeway, Gravitar, Montezuma's Revenge, Pitfall!, Private Eye, Solaris, and Venture。RL算法倾向于在这些游戏中挣扎，往往连一个正面的奖励都找不到。

特别是，《蒙特祖玛的复仇》被认为是RL代理的一个难题，需要结合掌握多种游戏中的技能来避免致命的障碍，并找到即使在最佳发挥下也相差数百步的奖励。通过获得专家示范（Pohlen等人，2018；Aytar等人，2018；Garmulewicz等人，2018）、对底层仿真器状态的特殊访问（Tang等人，2017；Stanton & Clune，2018）或两者（Salimans & Chen，2018）的方法已经取得了重大进展。然而，如果没有这些辅助工具，《蒙特祖玛的复仇》中的探索问题进展缓慢，最好的方法只能找到大约一半的房间（Bellemare等人，2016）。由于这些原因，我们在这个环境中对我们的方法进行了广泛的消融。

我们发现，即使完全不考虑外在奖励，最大化RND探索奖励的代理人也能持续找到蒙特祖马复仇记中一半以上的房间。为了将探索奖金与外在奖励结合起来，我们引入了近似政策优化（PPO，Schulman等人（2017））的修改，对两个奖励流使用两个价值头。这允许对不同的奖励使用不同的折现率，并结合偶发和非偶发的回报。有了这种额外的灵活性，我们最好的代理经常在蒙特苏马的复仇中找到第一层24个房间中的22个，并且偶尔（尽管不

经常）通过第一层。同样的方法在Venture和Gravitar上得到了最先进的表现。

2 方法

2.1 勘探红利

探索奖金是一类鼓励代理人探索的方法，即使在环境的奖励e_t是稀少的。它们通过用新的奖励$r_t = e_t + i_t$来取代e_t，其中i_t是与时间t的过渡相关的探索奖金。

为了鼓励代理人访问新的状态，希望i_t在新的状态中比经常访问的状态中更高。基于计数的探索方法提供了这种奖金的一个例子。在一个具有有限数量状态的表格设置中，我们可以将i_t定义为一个递减的函数
$_t$特别是$i_t = 1/n_t(\boldsymbol{s})$和$i_t = 1/n_t(\boldsymbol{s})$的访问数。
已在之前的工作中使用（Bellemare等人，2016；Ostrovski等人，2018）。在非表格的情况下，产生计数并不简单，因为大多数状态最多只被访问一次。计数在非表格环境中的一个可能的概括是伪计数（Bellemare等人，2016），它使用状态密度估计的变化作为探索奖励。这样一来，即使对于过去没有访问过的状态，只要它们与之前访问过的状态相似，从密度模型中得出的计数也可以是正数。

另一种方法是将i_t定义为与代理人的转换有关的问题的预测误差。这种问题的通用例子包括正向动力学（Schmidhuber，1991b；Stadie等人，2015；Achiam & Sastry，2017；Pathak等人，2017；Burda等人，2018）和逆向动力学（Haber等人，2018）。如果有关于环境的专门信息，也可以使用非通用的预测问题，比如预测代理人与之互动的物体的物理属性（Denil等人，2016）。这种预测误差往往会随着代理人收集更多与当前相似的经验而减少。由于这个原因，即使是微不足道的预测问题，如预测一个恒定的零函数，也能作为探索奖金发挥作用（Fox等人，2018）。

2.2 随机网络蒸馏

本文介绍了一种不同的方法，预测问题是随机产生的。这涉及到两个神经网络：一个是固定的、随机初始化的*目标*网络，它设定了预测问题，另一个是根据代理人收集的数据训练的*预测者*网络。目标网络将一个观察结果带到一个嵌入的f：$\mathcal{O} \rightarrow \mathbb{R}^k$，预测器神经网络$\hat{f}: \mathcal{O} \rightarrow \mathbb{R}^k$
通过梯度下降训练，使预期MSE最小化$\|\hat{f}(x;\theta) - f(x)\|^2$，关于其
参数$\theta_{\hat{f}}$。这个过程将一个随机初始化的神经网络提炼成一个经过训练的网络。对于与预测器训练过的状态不同的新状态，预计预测误差会更大。

为了建立直觉，我们考虑在MNIST上建立一个这个过程的玩具模型。我们在由标签为0的图像和目标类别的图像的混合物组成的训练数据上训练一个预测神经网络来模仿一个随机初始化的目标网络，改变类别的比例，但不改变训练例子的总数。然后，我们在未见过的目标类别的测试例子上测试预测网络，并报告MSE。在这个模型中，"零 "扮演的是以前多次出现的状态，而目标类扮演的是不常被访问的状态。结果显示在图2中。图中显示，测试误差随着目标类中训练例子数量的增加而减少，表明这种方法可以用来检测新奇性。图1显示，在《蒙特祖玛的复仇》的一集里，新奇状态下的内在奖励很高。

对这种方法的一个反对意见是，一个足够强大的优化算法可能会找到一个在任何输入上都能完美模仿目标随机网络的预测器（例如，目标网络本身就是这样一个预测器）。然而，上述关于MNIST的实验表明，基于梯度的标准方法不会以这种不可取的方式过度泛化。

2.2.1 预测误差的来源

一般来说，预测错误可以归因于一些因素：

1. *训练数据的数量*。在预测者看到的类似例子很少的情况下，预测误差很高（认识论的不确定性）。
2. *随机性*。预测误差很高，因为目标函数是随机的（alleatoric un- certainty）。随机过渡是前向动力学预测的这种误差的来源。
3. *模型的不规范性*。预测误差很大，因为缺少必要的信息，或者模型类别太有限，无法适应目标函数的复杂性。
4. *学习动态*。预测误差大是因为优化过程未能在模型类中找到一个最接近目标函数的预测器。

因素1是允许人们将预测错误作为探索的奖金。在实践中，预测误差是由所有这些因素的组合造成的，并不是所有的因素都是可取的。

例如，如果预测问题是前向动力学，那么因子2就会导致 "噪声-电视 "问题。这是一个思想实验，在这个实验中，一个因其前向动力学模型预测中的错误而得到奖励的代理人被环境中的局部熵源所吸引。一台显示白噪声的电视将是这样一个吸引者，就像一枚硬币的翻转一样。

为了避免不受欢迎的因素2和3，诸如Schmidhuber（1991a）；Oudeyer等人（2007）；Lopes等人（2012）；Achiam & Sastry（2017）的方法，而是使用测量预测模型在看到一个新的数据点时的改善程度。然而，这些方法往往在计算上很昂贵，因此难以扩展。

RND避免了因素2和3，因为目标网络可以被选择为确定性的，并且在预测网络的模型类别内。

2.2.2 与不确定性量化的关系

RND预测误差与Osband等人（2018）介绍的一种不确定性量化方法有关。也就是说，考虑一个回归问题，数据分布$D = \{x_i, y_i\}_{i}$。在贝叶斯设置中，我们会考虑一个关于映射f_θ□参数的先验$p(\theta^*)$，并在证据上更新后计算后验。

设F是函数$g_\theta = f_\theta + f_\theta$□的分布，其中$\theta^*$是从$p(\theta^*)$中抽取的，而$\theta$是通过最小化预期预测误差给出的。

$$\theta = \arg\min \mathsf{E}_{(x\,i,y_i)\square D}\| f_\theta(x_i) + f_\theta \square (x_i \} + \mathsf{R}(\theta)、 \quad (1)$$
$$- y_i \|_{\theta}$$

其中R（θ）是来自先验的正则化项（见Lemma 3，Osband等人（2018））。Osband等人（2018）认为（通过与贝叶斯线性回归的情况相类似），集合F是后验的近似值。

如果我们将回归目标y_i专门化为零，那么优化问题
$\arg_{min}\theta\, \mathsf{E}_{(x,y)\tilde{}D}\|f_\theta(x_i) + f_\theta \square (x)_i\|^2$ 相当于从随机抽取的函数中提炼出一个的先验。从这个角度看，预测器和目标网输出的每个坐标都对应于一个集合体的成员（集合体之间共享参数），而MSE将是对集合体预测方差的估计（假设集合体是无偏的）。换句话说，蒸馏误差可以被看作是对预测恒定零函数的不确定性的量化。

2.3 结合内在的和外在的回报

在只使用内在奖励的初步实验中，将问题视为非周期性的结果是更好的探索。在这种情况下，回报不会在 "游戏结束 "时被截断。我们认为这是在模拟环境中进行探索的自然方式，因为代理人的内在回报应该与它在未来可能发现的所有新的状态有关，不管它们是在一个情节中发生还是分散在几个情节中。Burda等人，2018）中也认为，使用情节性的内在奖励可以将任务的信息泄露给代理人。

我们还认为，这更接近于人类探索游戏的方式。例如，假设爱丽丝在玩一个电子游戏，并试图通过一个棘手的操作来到达一个可疑的秘密房间。因为这个动作很棘手，游戏结束的

几率很高，但如果爱丽丝成功了，她的好奇心的回报也会很高。如果爱丽丝被模拟成一个偶发的强化学习代理，那么，如果她在游戏中失败，她的未来回报将正好是零，这可能会使她过度厌恶风险。对爱丽丝来说，游戏结束的真正成本是不得不从头玩一遍游戏所产生的机会成本（估计爱丽丝在玩了一段时间的游戏后，对游戏的兴趣会降低）。

然而，使用非周期性回报的外在奖励可以被一种策略所利用，即在接近游戏开始时找到奖励，故意通过获得游戏结束来重启游戏，并在无尽的循环中重复这样做。

如何估计非周期性的内在奖赏流的综合价值并不明显
i_t，而外在奖励的偶发流e_t。我们的解决方案是观察回报是线性的。

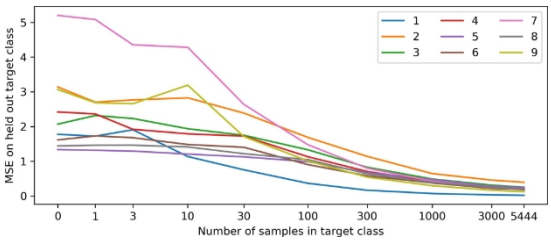


图2：MNIST上的新颖性检测：一个预测网络模仿随机初始化的目标网络。训练数据由不同比例的 "0 "类和目标类图像组成。每条曲线显示了对目标类例子的测试MSE，与目标类的训练例子的数量相对应（对数比例）。

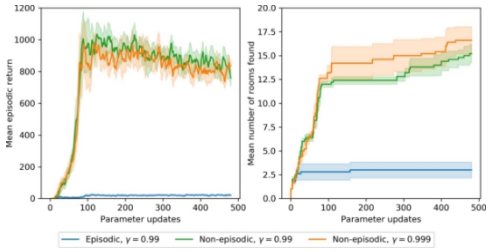


图3：在蒙特苏马的复仇游戏中，纯探索代理在没有获得外在奖励的情况下训练的平均偶发性返回和发现的房间数量。代理人在非偶发环境中探索得更多（另见第2.3节）。

奖励，因此可以分解为$R = R_E + R_I$，分别为外在和内在的回报之和。因此，我们可以用各自的回报分别拟合两个价值头V_E和V_I，并将它们结合起来，得到价值函数$V = V_E + V_I$。这个想法也可以用来结合具有不同折扣系数的回报流。

请注意，即使在不试图结合偶发和非偶发奖励流，或具有不同贴现因子的奖励流的情况下，具有独立的价值函数可能仍有好处，因为对价值函数有一个额外的监督信号。这对探索奖金可能特别重要，因为外在的奖励函数是静止的，而内在的奖励函数是非静止的。

2.4 奖励和观察的正常化

使用预测误差作为探索奖励的一个问题是，奖励的规模在不同的环境和不同的时间点上会有很大的不同，这使得我们很难选择在所有环境下都有效的超参数。为了使奖励保持在一个一致的尺度上，我们通过将内在奖励除以内在收益的标准偏差的运行估计值，将其规范化。

观测归一化在深度学习中通常很重要，但当使用随机神经网络作为目标时，它是至关重要的，因为参数是冻结的，因此不能根据不同数据集的规模进行调整。缺乏归一化会导致嵌入的方差极低，并携带关于输入的少量信息。为了解决这个问题，我们使用了连续控制问题中经常使用的观测归一化方案，即通过减去运行平均值，然后除以运行标准差，来增白每个维度。然后，我们将归一化的观测值剪切到-5和5之间。在开始优化之前，我们通过在环境中对一个随机代理进行少量的步骤来初始化归一化参数。我们对预测器和目标网络使用相同的观察值归一化，但不包括政策网络。

3 实验

在第3.1节中，我们首先在Montezuma's Revenge上进行了仅有内在奖励的实验，以隔离RND

奖金的归纳偏差，然后在第3.2-3.4节中对RND在Montezuma's Revenge上进行了广泛的消减，以了解促成RND性能的因素，最后在第3.6节中对6个硬探索Atari游戏与基线方法进行了对比。关于超参数和结构的细节，我们请读者参考附录A.3和A.4。大多数实验都是在每个环境下运行长度为128的30K次滚动，有128个平行环境，总共有30K次滚动。
19.7亿帧的经验。

3.1 纯粹的探索

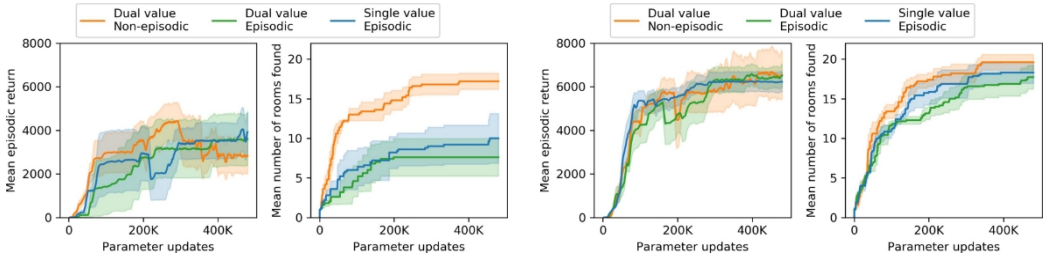
在这一节中，我们探讨了在没有任何外在奖励的情况下RND的性能。在本节中2.3 我们认为，在非偶发环境中，使用RND的探索可能更自然。通过比较纯勘探代理在偶发和非偶发环境中的性能，我们可以看到这一观察是否转化为改进的勘探性能。

我们在图3中报告了两个衡量探索性能的指标：平均偶发性回报，以及代理人在训练运行中发现的房间数量。由于纯粹的探索代理不知道外在的奖励或房间的数量，所以它没有直接优化任何这些措施。然而在蒙特苏马的复仇中获得一些奖励（如获得开门的钥匙）是获得新房间中更多有趣状态的必要条件，因此我们观察到外在奖励随着时间的推移而增加，直到某个点。当代理人与一些物体进行互动时，可以获得最好的回报，但一旦这种互动变得重复，代理人就没有动力继续做同样的事情，因此回报并不持续高。

我们在图3中清楚地看到，在这两种探索措施上，非偶发性代理表现最好，与第2.3节的讨论一致。$\gamma_I = 0.999$的非episodic设置比$\gamma_I = 0.99$的设置探索了更多的房间，其中一次运行探索了21个房间。在这种设置下，5次运行中的4次所取得的最佳回报是6700。

3.2 结合偶发和非偶发的回报

在第3.1节中，我们看到在没有任何外在奖励的情况下，非偶发环境比偶发环境导致更多的探索。接下来我们考虑在我们结合内在和外在奖励的情况下，这是否成立。正如第2.3节所讨论的，为了结合偶发和非偶发的奖励流，我们需要两个价值头。这也提出了一个问题：即使两个奖励流都是偶发的，是否有两个价值头更好。在图4中，我们比较了偶发的内在奖励和非偶发的内在奖励与偶发的外在奖励的结合，另外在偶发的情况下，两个价值头与一个价值头。折扣系数为$\gamma_I = \gamma_E = 0.99$。



(a) RNN政策　　(b) CNN政策

图4：结合内在奖励和外在奖励的不同方式。将非偶发的内在奖励流与偶发的外在奖励流结合起来，在探索的房间数量方面优于结合偶发版本的两种流，但在平均回报方面表现相似。综合外显回报流的单值估计比双值估计表现得更好一些。差异在RNN策略中更为明显。CNN的运行比RNN的同行更稳定。

在图4中我们看到，使用非偶发性的内在奖励流增加了CNN和RNN策略探索的房间数量，这

与第3.1节的实验一致，但差异不大，可能是因为外在奖励能够保留有用的行为。我们还看到，对于CNN的实验来说，差异不那么明显，RNN的结果往往不那么稳定，在$\gamma_E = 0.99$时表现更差。

与我们的预期相反（第2.3节），在偶发环境中，使用两个价值头并没有显示出比单一价值头有任何好处。然而，拥有两个价值头对于结合具有不同特征的奖励流是必要的，因此所有进一步的实验都使用两个价值头。

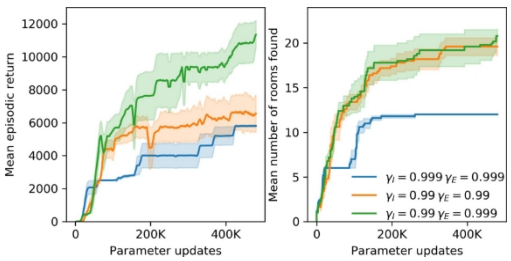


图5：不同折扣系数对内在和外在奖励流的表现。对外在奖励来说，较高的折扣系数会导致更好的表现，而对内在奖励来说，它损害了探索。

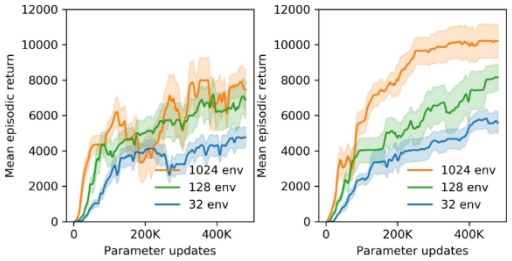


图6：对于CNN策略（左）和RNN策略（右）来说，随着用于收集经验的并行环境数量的增加，平均偶发性回报率也在提高。运行中处理了0.5,2和16B帧。

3.3 折扣因素

以前的实验（Salimans和Chen，2018；Pohlen等人，2018；Garmulewicz等人，2018）使用专家示范解决蒙特祖马的复仇，使用高折扣系数来实现
的最佳性能，使代理人能够预测到未来很长时间的奖励。我们比较了$\gamma_E \in \{0.99, 0.999\}$和$\gamma_I = 0.99$的RND代理的性能。我们还研究了将$\gamma_I$增加到0.999的效果。结果显示在图5中。

在图5中，我们看到将γ_E增加到0.999，而将γ_I保持在0.99，大大改善了性能。我们还看到，进一步增加γ_I到0.999会损害性能。这与图3的结果不一致，在图3中，增加γ_I并没有明显影响性能。

3.4 扩大培训规模

在这一节中，我们报告了显示增加规模对训练的影响的实验。本质奖励是非偶发性的，$\gamma_I = 0.99$，$\gamma_E = 0.999$。

为了在具有不同数量并行环境的实验中保持内在奖励随时间下降的速度不变，我们在训练预测器时降低了批次大小，以匹配32个并行环境的批次大小（全部细节见附录A.4）。较大的环境数会导致训练策略时每次更新的批量较大，而预测器网络的批量大小保持不变。由于内在奖励随着时间的推移而消失，因此政策必须学会寻找和利用这些短暂的奖励，因为它们是通往附近新状态的垫脚石。

图6显示，用从更多的平行环境中收集的较大批次的经验训练的代理在类似数量的更新后获得更高的平均回报。他们也取得了更好的最终性能。这种效果对于CNN策略来说似乎比RNN策略更早饱和。

我们允许有32个平行环境的RNN实验运行更多的时间，最终在处理了16亿帧的160万个参数更新后，达到了7570的平均回报。其中一次运行访问了所有24个房间，并通过了第一关，取得了17,500的最佳回报。有1024个平行环境的RNN实验在训练结束时，平均回报率为10,070，并产生了一个平均回报率为14,415的运行。

3.5 回归

Montezuma's Revenge是一个部分可观察的环境，尽管游戏状态的大部分可以从屏幕上推断出来。例如，代理拥有的钥匙数量出现在屏幕上，但不包括钥匙的来源，过去使用过多少钥匙，或者打开过哪些门。为了处理这种部分可观察性，一个代理应该保持一个总结过去的状态，例如一个经常性政策的状态。因此，很自然地希望具有经常性政策的代理人能有更好的表现。与预期相反，在图4中，递归政策的表现比非递归对应的政策差，$\gamma_E = 0.99$。然而，在图6中，RNN政策与

$\gamma_E = 0.999$，在每个尺度上都优于CNN的对应产品。[1].图7和图9的比较表明，在多个游戏中，RNN策略比CNN的表现更频繁。

3.6 与基线的比较

在本节中，我们将RND与两个基线进行比较：没有探索奖励的PPO和基于前向动力学误差的替代探索奖励。我们评估了RND在六个硬探索Atari游戏中的表现：Gravitar, Montezuma's Revenge, Pitfall!, Private Eye, Solaris, and Venture.我们首先与一个没有内在奖励的基线PPO实现的性能进行比较。对于RND来说，内在奖励是非偶发性的，$\gamma_I = 0.99$，而PPO和RND的$\gamma_E = 0.999$。图7显示了RNN策略的结果，并在表1中进行了总结（也见图9的CNN策略）。

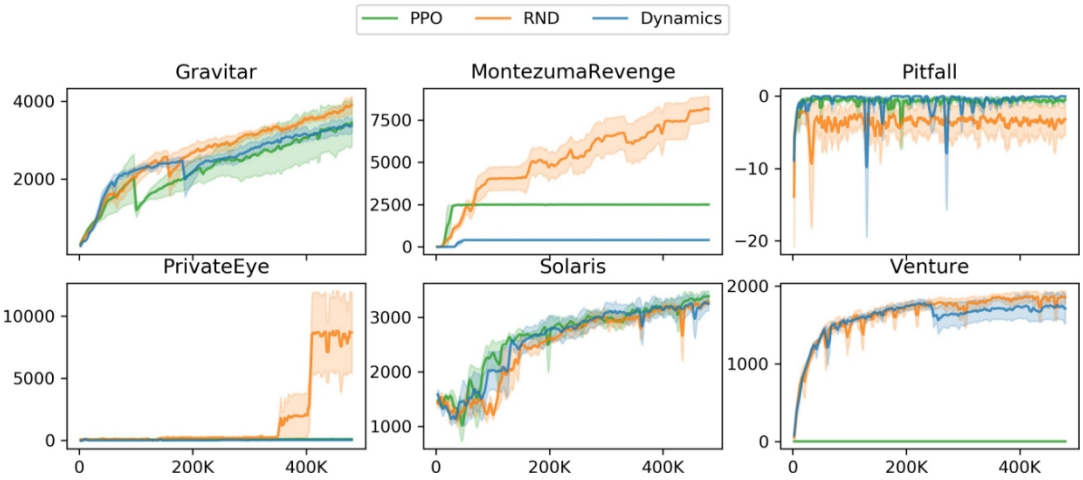


图7：基于RNN的政策的平均偶发性回报：RND、基于动力学的探索方法和仅有外在奖励的PPO在6个艰苦探索的Atari游戏上的表现。RND在Gravitar、Montezuma's Revenge和Venture上实现了最先进的性能，在后两者上明显优于PPO。

在Gravitar中，我们看到RND并没有持续超过PPO的性能。然而，两者都超过了使用RNN策略的人类平均性能，以及之前的技术水平。在Montezuma's Revenge和Venture中，RND明显地超过了PPO，并且超过了最先进的性能和人类的平均性能。在《Pitfall！》中，两种算法都未能找到任何积极的奖励。这是这个游戏的典型结果，因为外在的积极奖励是非常稀少的。在Private Eye上，RND的性能超过了PPO的性能。在Solaris上，RND的表现与PPO的表现相当。

接下来我们考虑基于前向动力学误差的另一种探索奖金。以前有许多作品使用这样的奖金（Schmidhuber，1991b；Stadie等人，2015；Achiam & Sastry，2017；Pathak等人，2017；Burda等人，2018）。幸运的是，Burda等人（2018）表明，在随机特征空间中训练前向动力学模型，当用于创建探索奖励时，通常与任何其他特征空间一样好用。这意味着我们可以很容易地实现苹果与苹果的比较，改变RND中的损失，以便预测器网络在当前观察和行动的情况下预测下一个观察的随机特征，同时保持我们方法的所有其他部分固定不变，如

双值头、非周期性的内在回报、规范化

[1]图5中CNN策略的结果是以5个随机种子的平均值获得的。当我们为图6的最佳表现设置运行10个不同的种子时，我们发现性能上有很大的差异。这种差异可能是由于 "蒙特苏马的复仇 "的结果分布以离散选择的影响为主（如从第一个房间向左或向右走），因此包含大量的离群值。此外，图5中的结果是用我们代码库的早期版本运行的，该版本与公开发布的版本之间的细微差别可能是造成差异的原因。图6中的结果是用公开发布的代码再现的，因此我们建议未来的工作与这些结果进行比较。

方案等。这提供了一个定义探索奖励的预测问题的消减，同时也是使用前向动力学误差的一类先前工作的代表。我们的期望是，除了基于动力学的代理能够利用环境中的非确定性来获得内在的奖励外，这些方法应该是相当相似的。

图7显示，在Montezuma's Revenge、PrivateEye和Solaris上，在相同的CNN策略下，基于动力学的探索表现明显不如RND，而在Venture、Pitfall和Gravitar上的表现相似。通过分析代理在收敛时的行为，我们注意到在Montezuma's Revenge中，代理在两个房间之间摇摆不定。这导致了不可减少的高预测误差，因为粘性行动的非确定性使得我们无法知道，一旦代理接近跨越房间边界，多走一步会导致它留在同一房间，还是跨越到下一个房间。这就是第2.2.1节中讨论的 "嘈杂的电视 "问题，或者说不确定性的表现。类似的行为也出现在PrivateEye和Pitfall中。表1中列出了每种算法的最终训练成绩，以及以前工作中的技术水平和人类平均成绩。

	重力器	蒙特祖马的复仇	坑爹啊!	私密的眼睛	阳光城	风险投资
RND	**3,906**	**8,152**	-3	8,666	3,282	**1,859**
PPO	3,426	2,497	0	105	3,387	0
动态性	3,371	400	0	33	3,246	1,712
神学院	2,209[1]	3,700[2]	**0**	**15,**[8062]	**12,**[3801]	**1,**[8133]
国家认可。人数	3,351	4,753	6,464	69,571	12,327	1,188

表1：与基线结果的比较。各种方法的最终平均性能。最先进的结果取自：[1] (Fortunato等人，2017) [2] (Bellemare等人，2016) [3] (Horgan等人，2018)

3.7 定性分析：与头骨共舞

通过观察RND代理，我们注意到，一旦它获得了它知道如何可靠地获得的所有外在奖励（由外在价值函数判断），该代理就会陷入一种行为模式，即不断与潜在的危险物体互动。例如，在《蒙特祖马的复仇》中，代理在一个移动的头骨上来回跳跃，在激光门之间移动，在消失的桥上来回走动。我们还在《坑爹》中观察到类似的行为。这可能与这样的事实有关，即这种危险的状态很难实现，因此与安全的状态相比，在代理人过去的经验中很少出现。

4 相关的工作

探索。基于计数的探索奖金是一种自然而有效的探索方式（Strehl & Littman, 2008），很多工作都研究了如何将计数奖金切实地推广到大的状态空间（Bellemare等人，2016；Fu等人，2017；Ostrovski等人，2017；Tang等人，2017；Machado等人，2018；Fox等人，2018）。

另一类探索方法依赖于预测动态的误差（Schmidhuber，1991b；Stadie等人，2015；Achiam

& Sastry，2017；Pathak等人，2017；Burda等人，2018）。正如第2.2节所讨论的，这些方法在随机或部分可观察的环境中会受到 "噪声电视 "问题的影响。这促使人们通过量化不确定性（Still & Precup，2012；Houthooft等人，2016）或预测改进措施（Schmidhuber，1991a；Oudeyer等人，2007；Lopes等人，2012；Achiam & Sastry，2017）进行探索。

其他探索方法包括对抗性自我博弈（Sukhbaatar等人，2018）、最大化授权（Gregor等人，2017）、参数噪音（Plappert等人，2017；Fortunato等人，2017）、识别多样化政策（Eysenbach等人，2018；Achiam等人，2018），以及使用价值函数的集合（Osband等人，2018；2016；Chen等人，2017）。

蒙特祖马的复仇。早期基于神经网络的强化学习算法在相当一部分雅达利游戏上取得了成功（Mnih等人，2015；2016；Hessel等人，2017），但却失败了

在蒙特祖马的复仇中取得有意义的进展，没有可靠地找到走出第一个房间的方法。这不一定是探索的失败，因为即使是一个随机的代理，每几十万步就能找到第一个房间的钥匙，每几百万步就能逃出第一个房间。事实上，在没有特殊探索方法的情况下，可以可靠地达到约2500的平均回报率（Horgan等人，2018；Espeholt等人，2018；Oh等人，2018）。

将DQN与伪数探索奖金相结合，Bellemare等人（2016年）创造了新的技术性能状态，探索了15个房间，得到了6600的最佳回报。此后，其他一些作品也取得了类似的性能（O'Donoghue等人，2017；Ostrovski等人，2018；Machado等人，2018；Osband等人，2018），没有超过它。

对底层RAM状态的特殊访问也可以用来改善探索，通过使用它来手工制作探索奖金（Kulkarni等人，2016；Tang等人，2017；Stanton & Clune，2018）。即使有这样的访问权，以前的工作取得的性能也不如人类的平均性能。

专家示范可以有效地用于简化《蒙特苏马的复仇》中的探索问题，一些作品（Salimans & Chen，2018；Pohlen等人，2018；Aytar等人，2018；Garmulewicz等人，2018）取得了与人类专家相当或更好的性能。从专家示范中学习得益于游戏的确定性。为了防止代理人简单地记住正确的动作序列，建议的训练方法（Machado等人，2017）是使用粘性动作（即随机重复以前的动作），但在这些工作中没有使用。在这项工作中，我们使用粘性动作，因此不依赖确定性。

随机特征。在监督学习的背景下，随机初始化神经网络的特征已被广泛研究（Rahimi & Recht，2008；Saxe等人，2011；Jarrett等人，2009；Yang等人，2015）。最近，它们被用于探索的背景中（Osband等人，2018；Burda等人，2018）。Osband等人（2018）的工作为第2.2节中讨论的随机网络提炼提供了动力。

矢量化的价值函数。Pong等人（2018）发现，矢量化的价值函数（其坐标对应于奖励的加法因素）改善了他们的方法。Bellemare等人（2017）将价值参数化为估计离散回报概率的价值头的线性组合。然而那里使用的贝尔曼备用方程本身并不是矢量的。

5 讨论

本文介绍了一种基于随机网络提炼的探索方法，并通过实验表明，该方法能够对几个奖励非常稀少的Atari游戏进行定向探索。这些实验表明，用相对简单的通用方法在困难的探索游戏上取得进展是可能的，特别是在大规模应用时。他们还表明，那些能够将内在奖励流与外在奖励流分开处理的方法（例如通过单独的价值头）可以从这种灵活性中受益。

我们发现，RND探索奖金足以处理局部探索，即探索短期决策的后果，如是否与特定对象互动，或避免它。然而，涉及长期协调决策的全球探索是我们的方法所不能及的。

要解决《蒙特祖马的复仇》的第一关，特工必须进入一个锁在两扇门后面的房间。整个关卡有四把钥匙和六扇门。四把钥匙中的任何一把都可以打开六扇门中的任何一扇，但在此过程中会被消耗掉。因此，为了打开最后两扇门，特工必须放弃打开两扇更容易找到的门，因为这两扇门打开后会立即得到奖励。

为了激励这种行为，代理人应该因保存钥匙而获得足够的内在奖励，以平衡早期使用钥匙所带来的内在奖励损失。从我们对RND代理行为的分析来看，它并没有得到足够大的激励来尝试这种策略，只是很少偶然发现它。

解决这个问题和需要高水平探索的类似问题是未来工作的一个重要方向。

参考文献

Joshua Achiam和Shankar Sastry。基于惊讶的内在动机的深度强化学习。 *arXiv:1703.01732*, 2017.

Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel.Variational option discovery algorithms. *arXiv preprint arXiv:1807.10299*, 2018.

Yusuf Aytar, Tobias Pfaff, David Budden, Tom Le Paine, Ziyu Wang, and Nando de Freitas.通过观看YouTube玩硬探索游戏。*arXiv预印本arXiv:1805.11592*，2018。

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos.统一基于计数的探索和内在动机。在*NIPS*, 2016.

Marc G Bellemare, Will Dabney, and Re´mi Munos.a distributionional perspective on reinforcement learning. *arXiv preprint arXiv:1707.06887*, 2017.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros.大规模的好奇心驱动的学习研究。In *arXiv:1808.04355*, 2018.

Richard Y Chen, John Schulman, Pieter Abbeel, and Szymon Sidor.通过*q-ensembles*的UCB和infogain探索。 *arXiv:1706.01502*, 2017.

Kyunghyun Cho, Bart Van Merrie¨nboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio.使用RNN编码器-解码器为统计机器翻译学习短语表征。*arXiv预印本arXiv:1406.1078*，2014。

Misha Denil, Pulkit Agrawal, Tejas D Kulkarni, Tom Erez, Peter Battaglia, and Nando de Fre- itas. 通过深度强化学习来学习进行物理实验。 *arXiv预印本arXiv:1611.01843*，2016。

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. IMPALA: Scalable distributed Deep-RL with importance weighted actor-learner architectures. *arXiv preprint arXiv: 1802.01561*, 2018.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine.Diversity is all you need： *arXiv preprint*, 2018.

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Ian Osband, Alex Graves, Vlad Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg.*arXiv:1706.10295*, 2017.

Lior Fox, Leshem Choshen, and Yonatan Loewenstein.探险家朵拉：定向外延式强化行动选择。*国际学习表征会议*，2018。

Justin Fu, John D Co-Reyes, and Sergey Levine.EX2：用典范模型探索深度强化学习。*NIPS*, 2017.

Michał Garmulewicz, Henryk Michalewski, and Piotr Miłos´.Expert-augmented actor-critic for vizdoom and montezumas revenge. *arXiv preprint arXiv:1809.03447*, 2018.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra.Variational intrinsic control.*ICLR研讨会*，2017。

Nick Haber, Damian Mrowca, Li Fei-Fei, and Daniel LK Yamins.*arXiv 预印本arXiv:1802.07442*, 2018.

Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver.彩虹：结合深度强化学习的改进。*arXiv预印本arXiv:1710.02298*，2017。

Dan Horgan, John Quan, David Budden, Gabriel Barth-Maron, Matteo Hessel, Hado Van Hasselt, and David Silver.分布式优先经验回放。*arXiv预印本arXiv:1803.00933*，2018。

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel.VIME：变量信息最大化的探索。在*NIPS*, 2016.

Kevin Jarrett, Koray Kavukcuoglu, Yann LeCun, 等. 什么是最好的物体识别多阶段结构？在*计算机视觉，2009年IEEE第12届国际会议上*，第2146-2153页。IEEE, 2009.

Diederik Kingma 和 Jimmy Ba.Adam: 一种随机优化的方法.*ICLR*, 2015.

Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum.层次化的深度强化学习：整合时间抽象和内在动机。In *Advances in neural information processing systems*, pp.3675-3683, 2016.

Manuel Lopes, Tobias Lang, Marc Toussaint, and Pierre-Yves Oudeyer.基于模型的强化学习中的探索，通过经验性地估计学习进度。在*NIPS*, 2012.

Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling.重新审视街机学习环境：*ArXiv preprint arXiv:1709.06009*, 2017.

Marlos C Machado, Marc G Bellemare, and Michael Bowling.*arXiv 预印本arXiv:1807.11622*, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller.*arXiv预印本arXiv:1312.5602*，2013年，用深度强化学习玩Atari。

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Belle- mare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis。通过深度强化学习实现人类层面的控制。*自然》*，518（7540）：529-533，2015年2月。

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu.深度强化学习的异步方法。在*ICML*，2016年。

Brendan O'Donoghue, Ian Osband, Remi Munos, and Volodymyr Mnih.不确定性贝尔曼方程和探索。*arXiv预印本arXiv:1709.05380*，2017。

Junhyuk Oh, Yijie Guo, Satinder Singh, and Honglak Lee.Self-imitation learning. *arXiv preprint arXiv:1806.05635*, 2018.

OpenAI.OpenAI五。https://blog.openai.com/openai-five/，2018。

OpenAI, :, M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki、A.Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba.学习灵巧的手部操作。*ArXiv e-prints*, August 2018.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy.通过引导的DQN进行深度探索。在*NIPS*, 2016.

Ian Osband, John Aslanides, and Albin Cassirer.用于深度强化学习的随机先验函数。 *arXiv预印本arXiv:1806.03335*，2018。

Georg Ostrovski, Marc G Bellemare, Aaron van den Oord, and Re´mi Munos.基于计数的探索与神经密度模型。 *arXiv:1703.01310*, 2017。

Georg Ostrovski, Marc G Bellemare, Aaron van den Oord, and Re´mi Munos.基于计数的神经密度模型的探索。*国际机器学习会议*，2018。

Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V Hafner.内在的动机系统，以促进自主精神的发展。*Evolutionary Computation*, 2007.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell.通过自我监督的预测进行好奇心驱动的探索。在*ICML*, 2017.

Matthias Plappert, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. 参数空间噪声的探索。*arXiv:1706.01905*，2017。

Tobias Pohlen, Bilal Piot, Todd Hester, Mohammad Gheshlaghi Azar, Dan Horgan, David Budden, Gabriel Barth-Maron, Hado van Hasselt, John Quan, Mel Vecˇer´ık, et al. Observe and look further：在Atari上实现一致的性能。*arXiv预印本arXiv:1805.11593*，2018。

Vitchyr Pong, Shixiang Gu, Murtaza Dalal, and Sergey Levine.时间差异模型：*arXiv preprint arXiv:1802.09081*, 2018.

Ali Rahimi和Benjamin Recht.大规模内核机的随机特征.In *Advances in neural information processing systems*, pp.1177-1184, 2008.

Tim Salimans 和 Richard Chen.从一次示范中学习蒙特祖马的复仇。*https://blog.openai.com/learning-montezumas-revenge-from-a-single-demonstration/*，2018年。

Andrew M Saxe, Pang Wei Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh, and Andrew Y Ng.关于随机权重和无监督的特征学习。In *ICML*, pp. 1089-1096, 2011.

Ju¨rgen Schmidhuber.好奇的模型构建控制系统。在*神经网络，1991年。1991年IEEE国际联合会议*，第1458-1463页。IEEE, 1991a.

Ju¨rgen Schmidhuber.在建立模型的神经控制器中实现好奇心和厌烦的可能性。*第一届适应性行为模拟国际会议论文集》*，1991b。

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov.近似的策略优化算法。*arXiv预印本arXiv:1707.06347*，2017。

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis。用深度神经网络和树状搜索掌握围棋游戏。*Nature*, 529(7587):484-489, Jan 2016.ISSN 0028-0836. doi: 10.1038/nature16961.

Bradly C Stadie, Sergey Levine, and Pieter Abbeel.用深度预测模型激励强化学习中的探索。*NIPS研讨会*，2015年。

Christopher Stanton 和 Jeff Clune.Deep curiosity search：生活中的探索提高了挑战性的深度强化学习问题的表现。 *arXiv预印本arXiv:1806.00553*, 2018.

Susanne Still和Doina Precup.信息理论方法用于好奇心驱动的强化学习.*生物科学理论》*，2012年。

Alexander L Strehl and Michael L Littman.基于模型的马尔科夫决策过程区间估计的分析.*Journal of Computer and System Sciences*, 74(8):1309-1331, 2008.

Sainbayar Sukhbaatar, Ilya Kostrikov, Arthur Szlam, and Rob Fergus.通过不对称的自我游戏的内在动机和自动课程。在*ICLR*, 2018.

唐浩然, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel.# 勘探：基于计数的深度强化学习探索的研究。In *NIPS*, 2017.

杨子超，Marcin Moczulski，Misha Denil，Nando de Freitas，Alex Smola，Le Song，和王紫玉。深度油炸的信念网。在*IEEE国际计算机视觉会议论文集中*，第1476-1483页，2015年。

Barret Zoph和Quoc V Le.*arXiv预印本arXiv:1611.01578*, 2016.

A 附录

A.1 强化学习算法

通过修改用于训练模型的奖励，探索奖励可以用于任何RL算法（即$r_t = i_t + e_t$）。我们将我们提出的探索奖金与基线强化学习算法PPO（Schulman等人，2017）相结合。PPO是一种策略梯度方法，我们发现它几乎不需要调整就能获得良好的性能。算法的细节见算法1。

A.2 rnd伪代码

算法1给出了RND方法的总体情况。该方法的确切细节可以在本文附带的代码中找到。

算法1 RND的伪代码

$N \leftarrow$ 轧制的数量
$N_{\text{opt}} \leftarrow$ 优化步骤的数量
$K \leftarrow$ 轧制的长度
$M \leftarrow$ 用于初始化观测规范化的初始步骤数
$t = 0$
样本状态 $s_0 \sim p_0(s)_0$
for $m = 1$ **to** M **do**
  样本 $a_t \sim \text{Uniform}(a_t)$ 样本
  $s_{t+1} \sim p(s_{t+1}|s_t, a)_t$
  使用s更新观测归一化参数$_{t+1}$
  t += 1
结束
for $i = 1$ **to** N **do**
  for $j = 1$ **to** K **do**
    样本 $a_t \sim \pi(a_t|s)_t$
    样本 $s_{t+1}, e_t \sim p(s_{t+1}, e_t|s_t, a)_t$
    计算内在奖励 $i_t = \|\hat{f}(s_{t+1}) - f(s)_{t+1}\|^2$
    将s_t, s_{t+1}, a_t, e_t, i_t 添加到优化批次B_i 使
    用i_t t += 1更新奖励标准化参数。
  结束
  将B中包含的内在奖赏归一化。$_i$
  计算收益$R_{I,i}$ 和内在奖励的优势$A_{I,i}$ 计算收益$R_{E,i}$ 和外在奖励的
  优势$A_{E,i}$ 计算综合优势$A_i = A_{I,i} + A_{E,i}$
  用B来更新观测归一化参数$_i$
  for $j = 1$ **to** N_{opt} **do**
    使用Adam优化θ_π 与PPO损失有关的批次 B_i, R_i, A_i 使用
    Adam优化$\theta_{\hat{f}}$与蒸馏损失有关的批次 B_i
  结束
结束

A.3 预处理细节

表2包含了我们如何为实验预处理环境的细节。我们遵循Machado等人（2017）的建议，使

用粘性动作，以使环境非决定性，这样就不可能记住动作序列了。在表3中，我们显示了政策和价值网络的额外预处理细节。在表4中，我们显示了预测器和目标网络的额外预处理细节。

超参数	价值
灰度缩放	True
观察下采样	(84,84)
外在奖励剪裁	[-1, 1]
内在奖励的削减	假的
每集最大帧数18K	丧失生命的终
端 虚假的最大和跳过的帧数4	
随机启动	False
粘性动作概率O	.25

表2：所有实验中环境的预处理细节。

超参数	价值
叠加的框架	4
观察	$x' \to x/255$
正常化	

表3：所有实验的政策和价值网络的预处理细节。

超参数	价值
叠加的框架	1
观察	$x' \to \text{CLIP}((x-\mu)/\sigma, [-5, 5])$。
正常化	

表4：所有实验的目标和预支配网络的预处理细节。

A.4 ppo和rnd超参数

在表5中，显示了PPO RL算法的超参数以及用于RND的任何附加超参数。关于如何使用这些超参数的完整细节可以在本文所附代码中找到。

超参数	价值
轧制长度	128
每个环境下的推广总数	30K
小型批次的数量	4
优化历时的数量	4
外在奖励的系数	2
内在奖赏的系数	1
并行环境的数量	128
学习率	0.0001
优化算法	亚当（Kingma & Ba (2015)）。
λ	0.95
熵系数	0.001
用于培训预测器的经验比例	0.25
γ_E	0.999
γ_I	0.99
夹子范围	[0.9, 1.1]
政策架构	有线电视新闻网

表5：PPO和RND算法的默认超参数，适用于实验。与这些默认值的任何差异都在正文中详细说明。

RND的最初初步实验只在32个平行环境下运行。我们预计，增加平行环境的数量将提高性

能，因为它允许政策更迅速地适应瞬息万变的内在奖赏。然而，如果预测器网络也能更快地学习，这种影响就会得到缓解。为了避免在从32个环境扩展到128个环境时出现这种情况，我们通过以0.25的概率随机丢弃批次中的元素来保持预测器网络的有效批次大小不变。同样，在256和1024个环境的实验中，我们分别以0.125和0.03125的概率放弃预测器的经验。

A.5 建筑

在本文中，我们使用了两个策略架构：一个RNN和一个CNN。两者都包含卷积编码器，与（Mnih等人，2015）中的标准架构相同。RNN架构还包含GRU（Cho等人，2014）单元以捕捉较长的上下文。政策的层大小被选择，以便参数的数量密切匹配。目标和预测网络的架构也有与（Mnih等人，2015）相同的卷积编码器，后面是密集层。确切的细节在本文附带的代码中给出。

A.6 其他实验结果

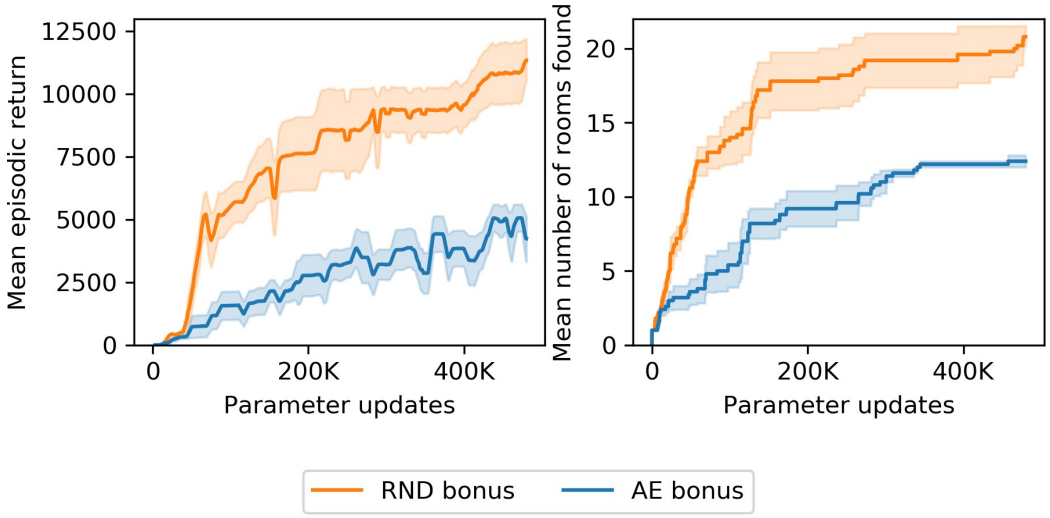


图8：RND与γ_I =0.99和γ_E =0.999的CNN策略的比较，探索由自动编码器的重建误差定义，保持所有其他选择不变（例如使用双值，将内在回报视为非episodic等）。基于自动编码器的代理的性能比RND差，但超过了基线PPO的性能。

图8比较了RND与相同算法的性能，但探索奖金定义为自动编码器的重建误差。自动编码任务在性质上与随机网络蒸馏相似，因为它也避免了2.2.1节中预测误差的第二个（尽管不一定是第三个）来源。实验表明，自动编码任务也可以成功用于探索。

图9比较了RND与PPO和基于动态预测的CNN策略基线的性能。

A.7 其他实验细节

在表6中，我们显示了每个实验所使用的种子数量，以图为索引。

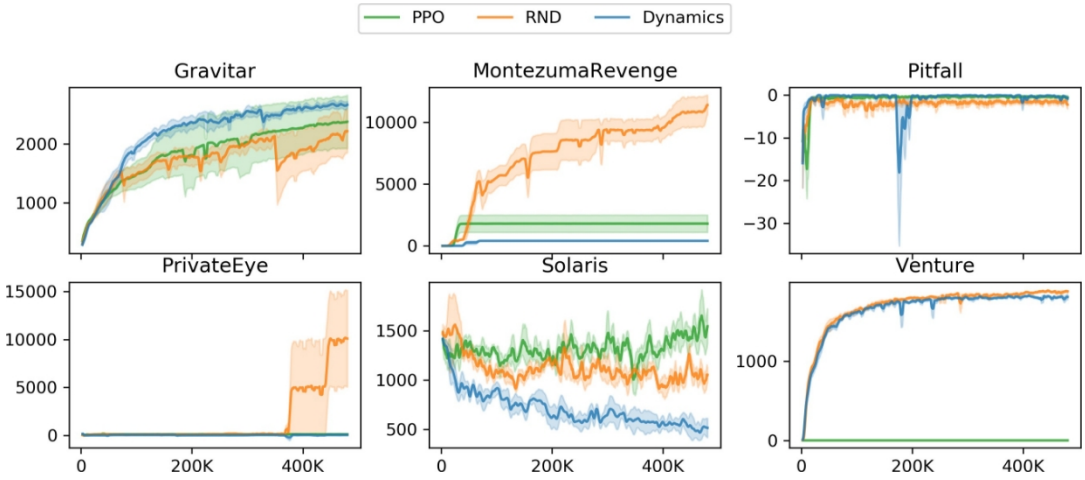


图9：基于CNN的政策的平均偶发性回报：RND、基于动力学的探索方法和仅有外在奖励的PPO在6个艰苦探索的Atari游戏上的平均回报。RND在Montezuma's Revenge、Private Eye和Venture上明显优于PPO。

	图号种子数量
1	这是对的。
2	10
3	5
4	5
5	10
6	5
7	3
8	5
9	5

表6：每个实验所运行的种子数量显示在表中。然后对每个种子的结果进行平均，在每个图中提供一个平均曲线，并使用标准误差使每个曲线周围的阴影区域。