

变压器的样品效率高

世界模型

文森特·米切利
日内瓦大学

Eloi Alonso
日内瓦大学

François Fleuret
日内瓦大学

摘要

深度强化学习代理是出了名的样本效率低下，这大大限制了它们在现实世界问题上的应用。最近，许多基于模型的方法被设计来解决这个问题，其中在世界模型的想象中学习是最突出的方法之一。然而，虽然与模拟环境的几乎无限的互动听起来很吸引人，但世界模型必须在很长一段时间内是准确的。在变形器在序列建模任务中成功的激励下，我们介绍了IRIS，一个在由离散自动编码器和自回归变形器组成的世界模型中学习的数据高效的代理。在Atari 100k基准测试中，IRIS只用了相当于两个小时的游戏时间，就获得了1.046的人类标准化平均分，并在26个游戏中的10个游戏中超过了人类，为没有前瞻搜索的方法设定了新的技术状态。为了促进未来对变形金刚和世界模型的研究，以提高样本效率的强化学习，我们在<https://github.com/eloialonso/iris>，发布了我们的代码和模型。

1 简介

深度强化学习（RL）已经成为在挑战性环境中开发有能力的代理的主导模式。最值得注意的是，深度RL算法在众多街机（Mnih等人，2015；Schrittwieser等人，2020；Hafner等人，2021）、实时策略（Vinyals等人，2019；Berner等人，2019）、棋盘（Silver等人，2016；2018；Schrittwieser等人，2020）和不完美信息（Schmid等人，2021；Brown等人，2020a）游戏中取得了令人瞩目的表现。然而，这些方法的一个共同缺点是其样本效率极低。事实上，经验要求从Atari 2600游戏（Bellemare等人，2013b）中DreamerV2（Hafner等人，2021）的几个月游戏时间到Dota2（Berner等人，2019）中OpenAI Five的几千年游戏时间。虽然有些环境可以为训练代理加速，但现实世界的应用往往不能。此外，可能会出现与环境互动的数量有关的额外成本或安全考虑（Yampolskiy，2018）。因此，样本效率是弥合研究和在野外部署深度RL代理之间差距的必要条件。

基于模型的方法（Sutton & Barto, 2018）构成了实现数据效率的一个有希望的方向。最近，世界模型以几种方式被利用：纯表示法学习（Schwarzer等人，2021）、前瞻搜索（Schrittwieser等人，2020；Ye等人，2021）和想象中的学习（Ha & Schmidhuber，2018；Kaiser等人，2020；Hafner等人，2020；2021）。后一种方法特别吸引人，因为在一个世界

模型内训练一个代理，使其摆脱了样本效率的限制。然而，这个框架在很大程度上依赖于准确的世界模型，因为政策纯粹是在想象中训练的。在一项开创性的工作中，Ha & Schmidhuber (2018) 成功地在玩具环境中建立了基于想象力的代理。SimPLe最近在更具挑战性的Atari 100k基准测试中显示了前景 (Kaiser等人, 2020)。目前，在想象力中学习的最好的Atari代理是DreamerV2 (Hafner等人, 2021)，尽管它是在有两亿帧可用的情况下开发和评估的，远远没有达到样本有效的制度。因此，设计新的世界模型架构，能够处理视觉上复杂和部分可观察的环境，而样本很少，是实现其作为训练场潜力的关键。

Transformer架构 (Vaswani等人, 2017) 现在在自然语言处理中无处不在 (Devlin等人, 2019; Radford等人, 2019; Brown等人, 2020b; Raffel等人, 2020)，并且在计算机视觉中也越来越受欢迎 (Dosovitskiy等人, 2021; He等人, 2022)，以及离线

*贡献相等，顺序由掷硬币决定。通信：{first.last}@unige.ch

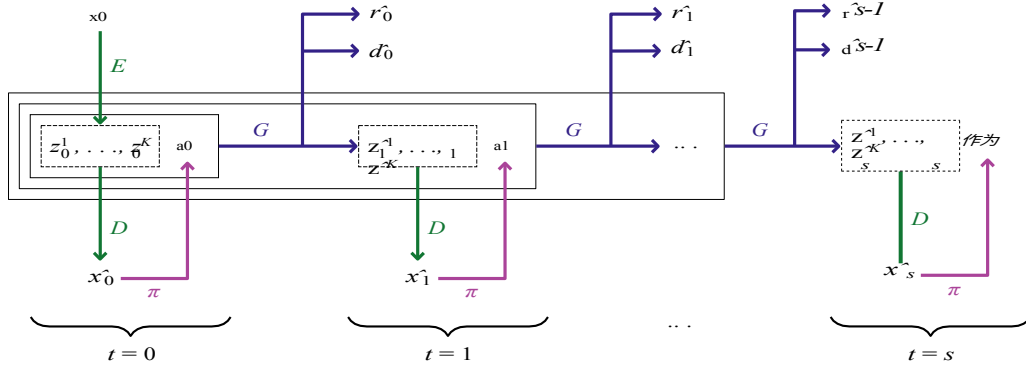


图1: 随着时间的推移展开想象力。该图显示了用紫色箭头描述的策略 π , 在想象中采取了一系列的行动。绿色箭头对应于离散自动编码器的编码器 E 和解码器 D , 其任务是用学到的符号语言表示框架。世界模型的主干 G 是一个类似GPT的转化器, 用蓝色箭头表示。对于策略 π 采取的每一个行动, G 通过自动渐进地展开 D 可以解码的新帧标记来模拟环境的动态。 G 还预测了奖励和潜在的情节终止。更具体地说, 一个初始帧 x_0 被 E 编码为标记 $z_0 = (z_0^1, \dots, z_0^K) = E(x_0)$ 。解码器 D 重建一个图像 $x_0^{\wedge} = D(z_0)$, 从这个图像中可以看到政策 π 预测行动 a_0 。从 z_0 和 a_0 , G 预测奖励 r_0^{\wedge} , 情节终止 $d_0^{\wedge} \in \{0, 1\}$, 并以自回归的方式 $z_1^{\wedge} = (z_1^{\wedge 1}, \dots, z_1^{\wedge K})$, 即下一帧的标记。A

虚线框表示给定时间步骤的图像标记, 而实线框表示 G 的输入序列, 即 $t=0$ 时的 (z_0, a_0) , $t=1$ 时的 $(z_0, a_0, z_1^{\wedge}, a_1)$, 等等。策略 π 纯粹是用想象的轨迹训练出来的, 只在真实环境中部署, 以改善世界模型 (E 、 D 、 G)。

强化学习 (Janner等人, 2021; Chen等人, 2021)。特别是GPT (Radford等人, 2018; 2019; Brown等人, 2020b) 系列模型在语言理解任务中取得了令人瞩目的成果。与世界模型类似, 这些基于注意力的模型是用高维信号和自我监督的学习目标来训练的, 因此构成了模拟环境的理想人选。

变换器在对离散标记的序列进行操作时尤其闪亮 (Devlin等人, 2019; Brown等人, 2020b)。对于文本数据, 有简单的方法 (Schuster & Nakajima, 2012; Kudo & Richardson, 2018) 来建立一个词汇表, 但是对于图像, 这种转换并不直接。一个天真的方法是将像素作为图像标记, 但标准的Transformer架构随着序列长度的增加而呈四次方扩展, 使得这个想法在计算上难以实现。为了解决这个问题, VQGAN (Esser等人, 2021) 和DALL-E (Ramesh等人, 2021) 采用离散自动编码器 (Van Den Oord等人, 2017) 作为从原始像素到更小数量的图像标记的映射。与自回归变换器相结合, 这些方法展示了强大的无条件和有条件的图像生成能力。这样的结果提出了一种设计世界模型的新方法。

在目前的工作中, 我们介绍了IRIS (通过内在语音自动回归的想象力), 这是一个在由离散自动编码器和自回归变换器组成的世界模型的想象中训练的代理。IRIS通过精确模拟数百万条轨迹来学习行为。我们的方法是将动态学习作为一个序列建模问题, 其中自动编码器建立了一种图像标记的语言, 而转化器则随着时间的推移构成了这种语言。通过最小的调整

，IRIS在Atari 100k基准（Kaiser等人，2020；Hessel等人，2018；Laskin等人，2020；Yarats等人，2021；Schwarzer等人，2021）的样本效率RL方面优于一系列最新方法（Kaiser等人，2020）。仅仅经过两个小时的实时体验，它就达到了1.046的平均人类归一化分数，并在26个游戏中的10个达到了超人的表现。我们在第2节描述了IRIS，并在第3节介绍了我们的结果。

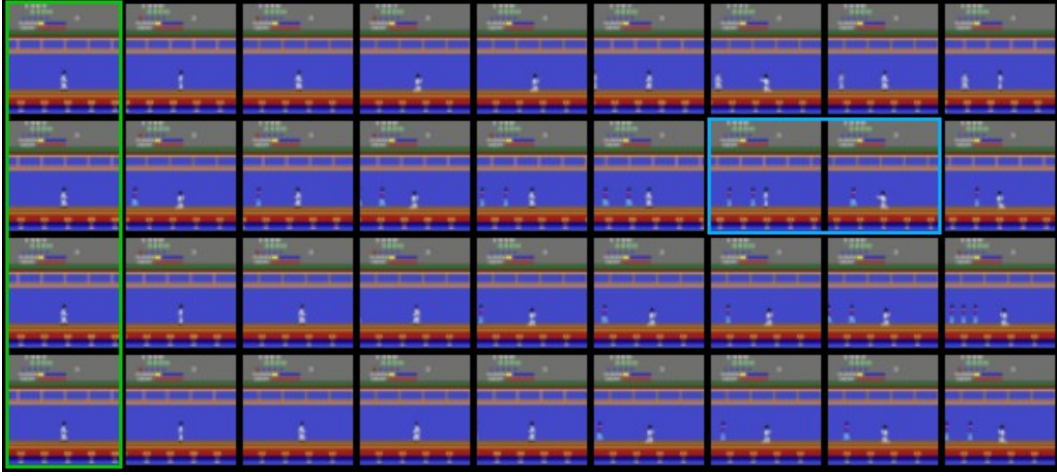


图2: *KungFuMaster*中的四条想象的轨迹。我们在四行中使用相同的调节帧，以绿色表示，其余的让世界模型来想象。由于初始帧只包含玩家，所以没有关于接下来的敌人的信息。因此，世界模型在每次模拟中都会产生不同类型和数量的对手。它能够反映出一个基本的游戏机制，在蓝色方框中突出显示，第一个敌人在被玩家击中后会消失。

2 方法

我们将问题表述为部分可观察马尔科夫决策过程（POMDP），图像观测值 $x_t \in \mathbb{R}^{h \times w \times 3}$ ，离散行动 $a_t \in \{1, \dots, A\}$ ，标量奖励 $r_t \in \mathbb{R}$ ，情节期限 $d_t \in \{0, 1\}$ ，折扣系数 $\gamma \in (0, 1)$ ，初始观测分布 ρ_0 ，以及环境动态 $x_{t+1}, r_t, d_t \sim p(x_{t+1}, r_t, d_t | x_{\leq t}, a_{\leq t})$ 。强化学习的目标是训练一个策略 π ，该策略产生的行动能使预期奖励之和最大化 $\mathbb{E}_{\pi} [\sum_{t \geq 0} \gamma^t r'_t]$ 。

我们的方法依赖于在想象中学习的三个标准组件（Sutton & Barto, 2018）：经验收集、世界模型学习和行为学习。按照 Ha & Schmidhuber（2018）；Kaiser等人（2020）；Hafner等人（2020；2021）的思路，我们的代理人完全在其世界模型中学习行动，我们只利用真实经验来学习环境动态。

我们反复进行以下三个步骤：

- `collect_experience`: 在真实环境中收集当前政策的经验。
- `update_world_model`: 改进奖励、情节结束和下一次观察的预测。
- `update_behavior`: 在想象中，改进政策和价值函数。

世界模型由一个离散自动编码器（Van Den Oord等人，2017）和一个类似GPT的自回归变换器（Vaswani等人，2017；Radford等人，2019；Brown等人，2020b）组成，其任务是捕捉环境的动态。图1说明了在想象过程中政策和这两个组件之间的相互作用。我们首先在第2.1和2.2节中分别描述了自动编码器和转化器。然后，第2.3节详细介绍了在想象中学习策略和价值函数的程序。附录A提供了对模型架构和超参数的全面描述。算法1总结了训练协议。

2.1 从图像观察到代币

离散自动编码器 (E, D) 学习了一种自己的符号语言，将高维图像表示为少量的标记。图1中的绿色箭头说明了帧和标记之间的来往。

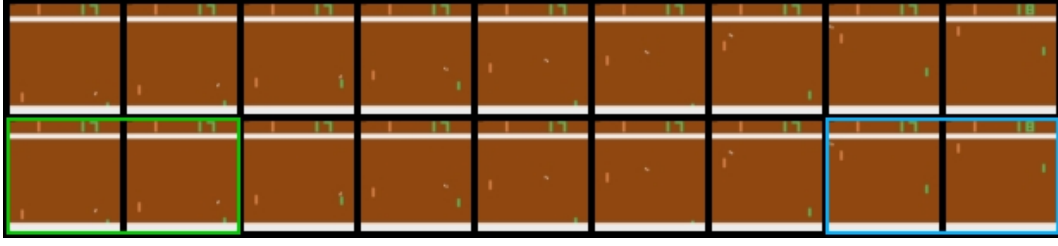


图3: *Pong*中的像素完美预测。上行显示的是在真实环境中收集的测试轨迹。底部一行描述了世界模型中该轨迹的重演。更确切地说,我们用绿色的真实序列的前两帧作为世界模型的条件。然后,我们依次给它提供真实的动作,让它想象随后的帧。仅仅经过120场训练,世界模型就完美地模拟了球的轨迹和球员的动作。值得注意的是,它还捕捉到了赢得交换后更新记分牌的游戏机制,如蓝色方框中所示。

更确切地说,编码器 $E: \mathbb{R}^{h \times w \times 3} \rightarrow \{1, \dots, N\}^K$ 将输入图像 x_t 转换为 K 从大小为 N 的词汇表中选取标记。让 $E = \{e_i\}_{i=1}^N \in \mathbb{R}^{N \times d}$ 是相应的嵌入 d 维向量的表格。输入图像 x_t 首先通过卷积神经网络 (CNN) (LeCun等人, 1989) 产生输出 $v_t \in \mathbb{R}^{K \times d}$ 。然后我们得到输出tokens $z_t = (z_t^1, \dots, z_t^K) \in \{1, \dots, N\}^K$, 因为 $z_t^k = \text{argmin}_i \|v_t^k - e_i\|$, 最接近的索引 i 的嵌入向量 (Van Den Oord等人, 2017; Esser等人, 2021)。反之, CNN解码器 $D: \{1, \dots, N\}^K \rightarrow \mathbb{R}^{h \times w \times 3}$ 将 K 个标记变回一个图像。

这个离散的自动编码器是在以前收集的帧上训练的, 有一个 L_1 重建损失、承诺损失 (Van Den Oord等人, 2017; Esser等人, 2021) 和感知损失 (Esser等人, 2021; Johnson等人, 2016; Larsen等人, 2016) 的等权组合。我们使用直通式估计器 (Bengio等人, 2013) 来实现反向传播训练。

2.2 建模动态

在高层次上, Transformer G 通过对离散自动编码器的语言进行建模来捕捉环境的动态变化。它的核心作用是展开想象力, 图1中的蓝色箭头强调了这一点。

具体来说, G 对交错的框架和动作标记的序列进行操作。一个输入序列 $(z^1, \dots, z^K, a_0, z^1, \dots, z^K, a_1, \dots, a_{t-1}, z^1, \dots, z^K, a_t)$ 是由原始序列得到的。

$(x_0, a_0, x_1, a_1, \dots, x_t, a_t)$ 通过对帧进行 E 编码, 如第2.1节所述。在每个时间

步长 t , 转化器对以下三个分布进行建模:

$$\text{过渡期: } \begin{aligned} z_{t+1}^k &\sim p_G(z_{t+1}^k | z_{\leq t}, a_{\leq t}) \text{ 与 } z_t^k \sim p_G(z_t^k | z_{\leq t}, a_{\leq t}, z_{t+1}^{\leq k}) \\ r_t &\sim p_G(r_t | z_{\leq t}, a_{\leq t}) \end{aligned} \quad (1)$$

$$r_t \sim p_G(r_t | z_{\leq t}, a_{\leq t}) \quad (2)$$

$$\text{奖励: } \begin{aligned} d_t &\sim p_G(d_t | z_{\leq t}, a_{\leq t}) \\ a_{\leq t} &\end{aligned} \quad (3)$$

请注意, 第 k 个标记的条件还包括 $z_{t+1}^{\leq k} := (z_{t+1}^1, \dots, z_{t+1}^k)$ 。这些标记是: $t+1$ 已被预测, 即自回归过程发生在符号层面。

我们以自我监督的方式在 L 个时间步骤的片段上训练 G , 这些片段是从过去的经验中抽取的。我们对过渡和终止预测器使用交叉熵损失, 对奖励预测器使用均方误差损失或交叉熵损

失，具体取决于奖励函数。

2.3 在想象中学习

离散自动编码器 (E, D) 和变换器 G 共同构成了一个世界模型，能够进行想象。图1中用紫色箭头描述的政策 π ，专门在这个想象力MDP中学习。

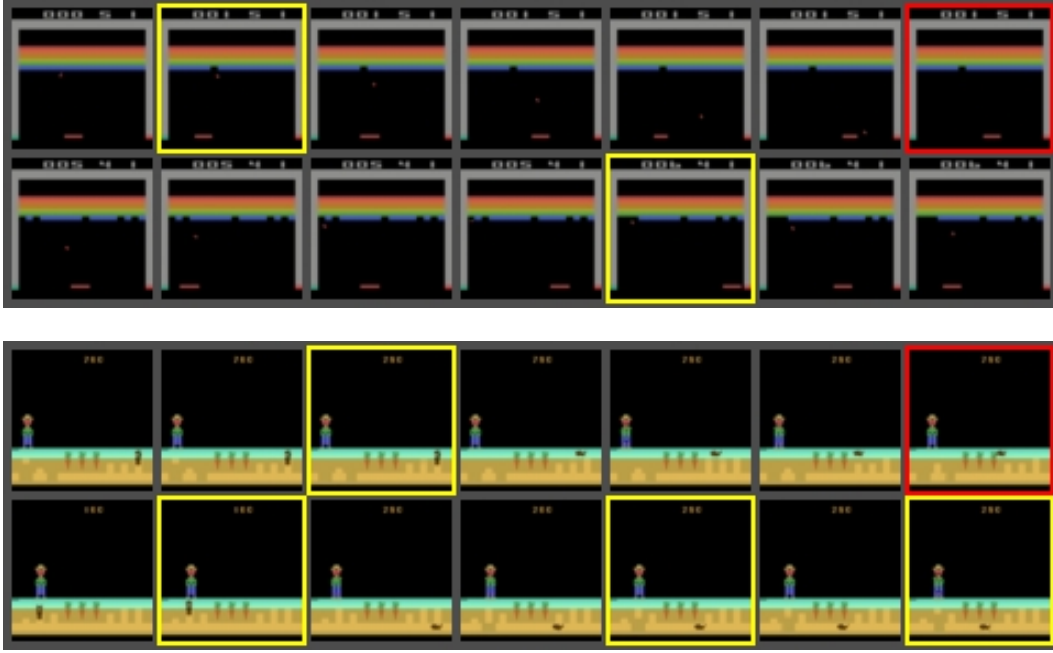


图4：在Breakout（上）和Gopher（下）中想象奖励和情节结束。每一行都描述了一个想象中的轨迹，并以真实环境中的一帧为初始化。黄框表示世界模型预测有正面奖励的帧。在Breakout中，它捕捉到打破砖头会产生奖励，并且砖头会在接下来的帧中被正确移除。在《打地鼠》中，玩家必须保护胡萝卜不被老鼠吃掉。世界模型成功地内化了堵住一个洞或杀死一个敌人会导致奖励。预测的情节终止用红框标出。世界模型准确地反映出，在《突围》中错过球，或在《打地鼠》中让敌人拿到胡萝卜，将导致一集的结束。

在时间步骤 t ，策略观察到一个重建的图像观察 \hat{x}_t ，并对行动 $a_t \sim \pi(a_t | \hat{x}_{\leq t})$ 进行采样。然后，世界模型预测奖励 \hat{r}_t ，情节结束 \hat{d}_t ，以及下一个观察 $\hat{x}_{t+1} = D(\hat{z}_{t+1})$ ，其中 $\hat{z}_{t+1} \sim p_G(\hat{z}_{t+1} | \hat{z}_0, a_0, \hat{z}_1, a_1, \dots, \hat{z}_t, a_t)$ 。这个想象程序以一个从过去经验中抽出的真实观测值 x_0 为初始化，并展开 H 步、

想象力水平线的超参数。如果在到达地平线之前预测到一个情节的结束，我们就停止。图1说明了想象的过程。

由于我们推出的想象力有固定的步数，我们不能简单地使用蒙特卡洛估计的预期回报。因此，为了引导代理人在给定的时间步骤之后会得到的回报，我们有一个价值网络 V ，它估计 $V(\hat{x}_t) = \mathbb{E}_{\pi, \tau \geq t} \gamma^{t-\tau} r_\tau$ 。

许多演员批评方法可以用来训练 π 和 V 的想象力（Sutton & Barto, 2018; Kaiser等人, 2020; Hafner等人, 2020）。为了简单起见，我们选择了DreamerV2（Hafner等人, 2021）的学习目标和超参数，该方法在Atari游戏中表现强劲。附录B给出了强化学习目标详细分类。

3 实验

采样效率高的强化学习是一个不断发展的领域，在复杂的视觉环境中有多基准（Hafner, 2022; Kanervisto等人, 2022）。在这项工作中，我们专注于成熟的Atari 100k基准（Kaiser等人, 2020）。我们在第3.1节中介绍了该基准和它的基线。我们在第3.2节中描述了评估协议并讨论了结果。第3.3节给出了世界模型能力的定性例子。

表1: 经过2小时的实时体验后, 雅达利100k的26个游戏的回报率, 以及人类规范化的综合指标。黑体数字表示没有前瞻搜索的顶级方法, 而下划线的数字则表示总体上最好的方法。IRIS在超人游戏的数量、平均值、四分位数 (IQM) 和最优性差距方面都优于单纯的学习方法。

游戏	随机	人类	前瞻性搜索		无提前量搜索				
			MuZero	效益零度	模拟 PLe	CURL	邓小平	纵横四海	IRIS (我们的)
外星人	227.8	7127.7	530.0	808.5	616.9	711.0	865.2	841.9	420.0
Amidar	5.8	1719.5	38.8	148.6	74.3	113.7	137.8	179.7	143.0
殴打	222.4	742.0	500.1	1263.1	527.2	500.9	579.6	565.6	1524.4
阿斯特里克斯	210.0	8503.3	1734.0	<u>25557.8</u>	1128.3	567.2	763.6	962.5	853.6
银行劫案	14.2	753.1	192.5	<u>351.0</u>	34.2	65.3	232.9	345.4	53.1
战斗地带	2360.0	37187.5	7687.5	13871.2	4031.2	8997.8	10165.3	14834.1	13074.0
拳击	0.1	12.1	15.1	52.7	7.8	0.9	9.0	35.7	70.1
突围赛	1.7	30.5	48.0	<u>414.1</u>	16.4	2.6	19.8	19.6	83.7
斩波器命令	811.0	7387.8	1350.0	1117.3	979.4	783.5	844.6	946.3	1565.0
疯抢者	10780.5	35829.4	56937.0	<u>83940.2</u>	62583.6	9154.4	21539.0	36700.5	59324.2
恶魔的攻击	152.1	1971.0	3527.0	<u>13003.9</u>	208.1	646.5	1321.5	517.6	2034.4
高速公路	0.0	29.6	21.8	21.8	16.7	28.3	20.3	19.3	31.1
冻伤	65.2	4334.7	255.0	296.3	236.9	1226.5	1014.2	1170.7	259.1
地鼠	257.6	2412.5	1256.0	<u>3260.3</u>	596.8	400.9	621.6	660.6	2236.1
英雄	1027.0	30826.4	3095.0	<u>9315.9</u>	2656.6	4987.7	4167.9	5858.6	7037.4
詹姆斯邦德	29.0	302.8	87.5	<u>517.0</u>	100.5	331.0	349.1	366.5	462.7
袋鼠	52.0	3035.0	62.5	724.1	51.2	740.2	1088.4	3617.4	838.2
克鲁尔	1598.0	2665.5	4890.8	5663.3	2204.8	3049.2	4402.1	3681.6	6616.4
巩俐老师	258.5	22736.3	18813.0	<u>30944.8</u>	14862.5	8155.6	11467.4	14783.2	21759.8
MsPacman	307.3	6951.6	1265.6	1281.2	1480.0	1064.0	1218.1	1318.4	999.1
Pong	-20.7	14.6	-6.7	<u>20.1</u>	12.8	-18.5	-9.1	-5.4	14.6
私密的眼睛	24.9	69571.3	56.3	96.7	35.0	81.9	3.5	86.0	100.0
Qbert	163.9	13455.0	3952.0	<u>13781.9</u>	1288.8	727.0	1810.7	866.3	745.7
路人甲	11.5	7845.0	2500.0	<u>17751.3</u>	5640.6	5006.1	11211.4	12213.1	9614.6
海底捞	68.4	42054.7	208.0	<u>1100.2</u>	683.3	315.2	352.3	558.1	661.3
向上向下	533.4	11693.2	2896.9	<u>17264.2</u>	3350.3	2646.4	4324.5	10859.2	3546.2
#超级人类(个)	0	不适用	5	14	1	2	3	6	10
平均值(个)	0.000	1.000	0.562	<u>1.943</u>	0.332	0.261	0.465	0.616	1.046
中位数(个)	0.000	1.000	0.227	<u>1.090</u>	0.134	0.092	0.313	0.396	0.289
IQM(个)	0.000	1.000	不适用	不适用	0.130	0.113	0.280	0.337	0.501
最优性差距 (↓)	1.000	0.000	不适用	不适用	0.729	0.768	0.631	0.577	0.512

3.1 基准和基线

Atari 100k由26个Atari游戏组成 (Bellemare等人, 2013a), 具有各种机制, 评估了广泛的代理能力。在这个基准中, 一个代理在每个环境中只允许有100k的行动。这一约束大致相当于人类游戏的2小时。作为比较, 无约束的Atari代理通常被训练了5000万步, 经验增加了500倍。

在Atari 100k基准上比较了多个基线。SimPLe (Kaiser等人, 2020) 在视频生成模型中用PPO (Schulman等人, 2017) 训练一个政策。CURL (Laskin等人, 2020) 从用对比学习获得的高级图像特征中开发出非政策代理。DrQ (Yarats等人, 2021年) 增强了输入图像, 并在几次转换中平均了Q值估计。SPR (Schwarzer等人, 2021) 在增强的视图和相邻的时间步骤中强制执行输入图像的一致表示。上述基线带有额外的技术来提高性能, 如优先的经验重放 (Schaul等人, 2016), epsilon-greedy调度, 或数据增强。

我们对有前瞻搜索和无前瞻搜索的方法进行了区分。事实上, 依靠决策时搜索的算法 (

Silver等人, 2016; 2018; Schrittwieser等人, 2020) 可以极大地提高代理的性能, 但它们在计算资源和代码复杂性方面是有代价的。MuZero (Schrittwieser 等人, 2020) 和 EfficientZero (Ye等人, 2021) 是目前Atari 100k中基于搜索方法的标准。MuZero利用蒙特卡洛树搜索 (MCTS) (Kocsis & Szepesvári, 2006; Coulom, 2007) 作为策略改进运算符, 通过在世界模型的潜在空间中展开多个假设轨迹。EfficientZero在MuZero的基础上进行了改进, 引入了自我监督的一致性损失, 一次性预测短期内的收益, 并利用其世界模型修正非政策轨迹。

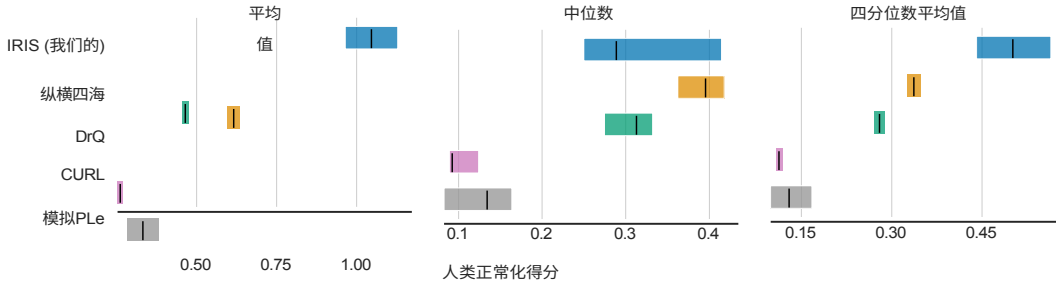
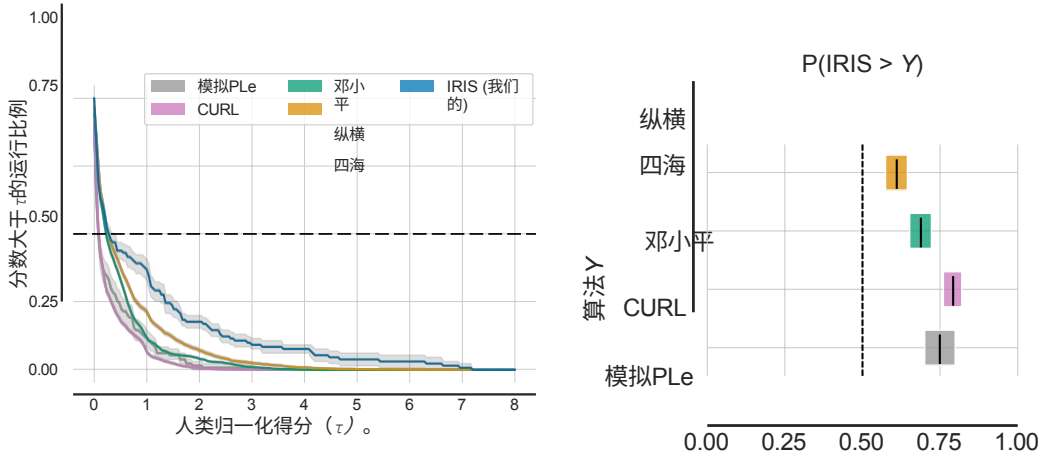


图5：平均数、中位数和四分位数的人类归一化分数，用分层的引导置信区间计算。IRIS和SimPLe为5次，SPR、CURL和DrQ为100次（Agarwal等人，2021）。



(a) 性能概况，即超过给定人类归一化分数的运行比例。(b) 改进的概率，即IRIS在任何游戏上超过基线的可能性有多大。

图6：性能概况（左）和改进的概率（右）（Agarwal等人，2021年）。

3.2 结果

人类正常化的分数是衡量Atari 100k性能的既定标准。它被定义为

$\frac{\text{score_agent} - \text{score_随机}}{\text{score_human} - \text{score_random}}$ ，其中 score_random 来自随机策略， score_human 来自人类玩家（Wang等人，2016）。

表1显示了不同游戏的收益和人类规范化的总指标。对于MuZero和EfficientZero，我们报告了Ye等人（2021）发表的平均结果（3次运行）。对于其他基线，我们使用Agarwal等人（2021）进行的Atari 100k案例研究的结果（CURL、DrQ、SPR的100次新运行和SimPLe的5次现有运行）。最后，我们通过计算每个游戏训练结束时收集的100个情节的平均值（5次运行）来评估IRIS。

Agarwal等人（2021）讨论了平均分和中位数的局限性，并表明在RL基准中，标准点估计值

和区间估计值之间会产生很大的差异。根据他们的建议，我们在图5中总结了人类归一化的分数，以及平均数、中位数和四分位数（IQM）的分层引导置信区间。为了进行更精细的比较，我们还在图6中提供了性能概况和改进的概率。

仅用相当于两小时的游戏时间，IRIS就获得了1.046的超人平均分（+70%），0.501的IQM（+49%），0.512的优化差距（+11%），并在26个游戏中的10个游戏中超过了人类玩家（+67%），其中的相对改进是相对于SPR计算的（Schwarzer等人，2021）。这些结果构成了在Atari 100k基准中没有前瞻搜索的方法的一个新的技术水平。我们还注意到，IRIS优于MuZero，尽管后者不是为样本效率制度设计的。



图7: *Frostbite* (左) 和 *Krull* (右) 游戏中的三个连续关卡。在我们的实验中, 世界模型在模拟 *Frostbite* 的后续关卡时很吃力, 但在 *Krull* 中则不然。事实上, 在《冰霜》中退出第一关需要一长串不太可能的动作, 首先要建造冰屋, 然后再从屏幕的底部回到冰屋。这种罕见的事件使世界模型无法内化游戏的新内容, 因此不会在想象中被政策所体验。虽然 *Krull* 的特点是关卡更加多样化, 但世界模型成功地反映了这种多样性, IRIS 甚至在这个环境中设定了一个新的艺术状态。这可能是由于在 *Krull* 中更频繁地从一个阶段过渡到下一个阶段, 导致对每个关卡的充分覆盖。

此外, 性能曲线 (图6a) 显示, IRIS 在其底部50%的游戏中与最强的基线持平, 在这一点上它随机地支配了 (Agarwal等人, 2021; Dror等人, 2019) 其他方法。同样地, 所有基线的改进概率都大于0.5 (图6b)。

在中位数得分方面, IRIS 与其他方法有重叠 (图5)。有趣的是, Schwarzer等人 (2021年) 指出, 中位数只受到少数决定性棋局的影响, 这一点从中位数得分的置信区间宽度可以看出, 即使 DrQ、CURL 和 SPR 运行了100次。

我们观察到, IRIS 在那些随着训练的进行不会出现分布性变化的游戏中表现得特别强。这类游戏的例子包括 *Pong*, *Breakout*, 和 *Boxing*。相反, 当一个新的关卡或游戏机制通过一个不可能的事件被解锁时, 代理就会很挣扎。这揭示了一个双重探索问题。IRIS 必须首先发现游戏的新方面, 使其世界模型内部化。只有这样, 政策才能重新发现并利用它。图7详细介绍了《冰霜之城》和《克鲁尔》这两款多关卡游戏的这一现象。总之, 只要关卡之间的转换不依赖于低概率事件, 双重探索问题就不会妨碍性能。

另一种难以模拟的游戏是视觉上具有挑战性的环境, 在这种环境中捕捉小的细节很重要。正如附录E中所讨论的, 增加编码帧的代币数量可以提高性能, 尽管是以增加计算量为代价的。

3.3 世界模型分析

由于 IRIS 完全在其想象中学习行为, 世界模型的质量是我们方法的基石。例如, 离散自动编

码器正确地重建球、球员或敌人等元素是关键。同样，变形器可能无法捕捉到重要的游戏机制，如奖励归属或情节终止，会严重妨碍代理人的表现。因此，无论有多少想象中的轨迹，如果世界模型有缺陷，代理将学习次优政策。

虽然第3.2节提供了定量评价，但我们的目的是用世界模型的能力的定性例子来补充分析。图2显示了在面临不确定性的情况下产生许多合理的未来的情况。图3描述了 *Pong* 中的像素完美预测。最后，我们在图4中说明了对奖励和情节终止的预测，这对强化学习的目标至关重要。

4 相关的工作

在世界模型的想象中学习

在学习到的世界模型中训练策略的想法首先在表格环境中得到了研究（Sutton & Barto, 2018）。Ha & Schmidhuber（2018）表明，简单的视觉环境可以用自动编码器和递归网络来模拟。SimPLe（Kaiser等人，2020）表明，在视频预测模型中训练的PPO政策（Schulman等人，2017）在一些Atari游戏中的表现优于人类。在Dreamer（Hafner等人，2020）的基础上进行改进，DreamerV2（Hafner等人，2021）是第一个在Atari 50M基准中达到人类水平的代理学习想象力。它的世界模型结合了卷积自动编码器和递归状态空间模型（RSSM）（Hafner等人，2019），用于潜伏动态学习。最近，Chen等人（2022）探索了DreamerV2的一个变体，其中一个Transformer取代了RSSM中的递归网络，Seo等人（2022）在有离线视频数据集可供预训练的情况下加强了DreamerV2。

用变形金刚进行强化学习

在自然语言处理方面取得惊人进展后（Manning & Goldie, 2022），强化学习社区最近也踏入了变形金刚的领域。Parisotto等人（2020）观察到，标准的变形器架构很难用RL目标进行优化。作者建议用门控层取代残余连接，以稳定学习程序。我们的世界模型不需要这样的修改，这很可能是由于其自我监督的学习目标。轨迹转化器（Janner等人，2021）和决策转化器（Chen等人，2021）将离线轨迹表示为序列的静态数据集，而在线决策转化器（Zheng等人，2022）将后者扩展到在线设置。轨迹转化器被训练来预测未来的回报、状态和行动。在推理时间，它可以通过奖励驱动的波束搜索来规划最佳行动，然而该方法仅限于低维状态。相反，决策转换器可以处理图像输入，但不容易扩展为世界模型。Ozair等人（2021）介绍了MuZero（Schrittwieser等人，2020）的一个离线变体，能够处理随机环境，通过对行动和轨迹级离散潜变量进行变形器混合搜索。

用离散的自动编码器和变压器生成视频

VQGAN（Esser等人，2021）和DALL-E（Ramesh等人，2021）使用离散自动编码器将一个帧压缩成一个小的标记序列，然后转化器可以对其进行自回归建模。其他工作将该方法扩展到视频生成。GODIVA（Wu等人，2021）为文本条件视频生成的帧序列而不是单一的帧建模。VideoGPT（Yan等人，2021）引入了视频级离散自动编码器，以及具有空间和时间注意力模式的变形器，用于无条件和行动条件的视频生成。

5 结论

我们介绍了IRIS，一个纯粹在由离散自动编码器和自回归变换器组成的世界模型的想象中学习的代理。IRIS在Atari 100k基准中为无前瞻搜索的方法设定了一个新的技术状态。我们表

明，其世界模型对游戏机制有深刻的理解，从而在一些游戏中实现了像素级的完美预测。我们还说明了世界模型的生成能力，在想象力训练时提供了丰富的游戏体验。最终，与现有的战斗力强的代理相比，IRIS以最小的调整开辟了一条有效解决复杂环境问题的新途径。

在未来，IRIS可以扩大到对计算要求高且具有挑战性的任务，这些任务将从其世界模型的速度中受益。此外，它的策略目前是从重建的框架中学习，但它可能会利用世界模型的内部表示。另一个令人兴奋的研究途径是将想象中的学习与MCTS相结合。事实上，这两种方法都能带来令人印象深刻的结果，它们对代理性能的贡献可能是互补的。

可重复性声明

第2节和附录B中介绍了不同的组件和它们的训练目标。我们在附录A中描述了模型的结构并列出了超参数。我们在附录G中说明了用于产生我们结果的资源。算法1明确了训练循环中各组成部分之间的相互作用。在第3.2节中，我们提供了基线的报告结果的来源，以及评估协议。

该代码是开源的，以确保结果的可重复性并促进未来的研究。运行该代码库需要最小的依赖性，我们提供了一份详尽的用户指南来开始使用。训练和评估可以通过简单的命令来启动，通过配置文件可以进行定制，我们还包括将代理人的游戏可视化并让用户与世界模型互动的脚本。

伦理声明

为现实世界环境开发自主代理引起了许多安全和环境问题。在训练期间，代理可能会对个人造成严重伤害，并破坏其周围环境。我们相信，在世界模型的想象中学习可以大大减少与训练新的自主代理有关的风险。事实上，在这项工作中，我们提出了一个世界模型架构，能够用很少的样本对环境进行精确建模。然而，在未来的研究路线中，人们可以更进一步，利用现有的数据来消除与真实世界互动的必要性。

鸣谢

我们要感谢Maxim Peter, Bálint Máté, Daniele Paliotta, Atul Sinha, 和Alexandre Dupuis的有见地的讨论和评论。Vincent Micheli得到了瑞士国家科学基金会的资助，资助号为FNS-187494。

参考文献

- Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. 在统计悬崖边缘的深度强化学习。 *Advances in neural information processing systems*, 34:29304-29320, 2021.
- Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. 拱廊学习环境：一个通用代理的评估平台。 *Journal of Artificial Intelligence Research*, 47: 253-279, 2013a.
- Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. 拱廊学习环境：一个通用代理的评估平台。 *Journal of Artificial Intelligence Research*, 47: 253-279, 2013b.
- Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 通过随机神经元估计或传播梯度的条件计算。 *arXiv预印本arXiv:1308.3432*, 2013。
- Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.
- Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong. 结合深度强化学习和搜索的不完美信息游戏。 *Advances in neural information processing systems*, 33:17057-17069, 2020a.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *神经信息处理系统的进展*, 33: 1877-1901, 2020b.
- Chang Chen, Yi-Fu Wu, Jaesik Yoon, and Sungjin Ahn. Transdreamer : *arXiv preprint arXiv:2202.09481*, 2022.
- Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 决策转化器：通过序列建模的强化学习。 *神经信息处理系统的进展*, 34, 2021。
- Rémi Coulom. 计算围棋游戏中棋型的 "埃洛等级"。 *ICGA杂志*, 30(4): 198-208, 2007.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT：用于语言理解的深度双向变换器的预训练。 见《*计算语言学协会北美分会2019年会议论文集*》：人类语言技术，第一卷（长篇和短篇论文），2019年。
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words：规模化的图像识别的变形器。在*国际学习表征会议上*，2021年。
- Rotem Dror, Segev Shlomov, and Roi Reichart. 深度优势-如何正确比较深度神经模型。 In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp.2773-2785, 2019.
- Patrick Esser, Robin Rombach, and Bjorn Ommer. 驯服变压器的高分辨率图像合成。 In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.12873-12883, 2021.

F.A. Gers, J. Schmidhuber, and F. Cummins. 学习遗忘：用LSTM进行连续预测。
神经计算, 12 (10) : 2451-2471, 2000。

David Ha 和 Jürgen Schmidhuber. 循环世界模型促进了政策演变。 *Advances in neural information processing systems*, 31, 2018.

Danijar Hafner.为代理人的能力谱系制定基准。在*国际学习代表会议上*，2022年。

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson.从像素中学习规划的潜在动力。在*国际机器学习会议上*，第2555-2565页。PMLR, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi.从梦想到控制：通过潜在的想象力学习行为。在*国际学习代表会议上*，2020年。

Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, and Jimmy Ba.用离散的世界模型掌握Atari。在*国际学习代表会议上*，2021年。

何开明，陈新磊，谢赛宁，李阳浩，Piotr Dollár，和Ross Girshick。遮蔽的自动编码器是可扩展的视觉学习器。在*IEEE/CVF 计算机视觉和模式识别会议论文集*中，第16000-16009页，2022年。

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。在*第三十二届AAAI人工智能会议上*，2018。

Sepp Hochreiter和Jürgen Schmidhuber.长短期记忆。*神经计算*, 9(8): 1735-1780, 1997.

Michael Janner, Qiyang Li, and Sergey Levine.离线强化学习是一个大的序列建模问题。*神经信息处理系统的进展*, 34, 2021.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei.用于实时风格转移和超级分辨率的感知损失。在*欧洲计算机视觉会议上*，第694-711页。Springer, 2016.

Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Miłoś, Błażej Osin'ski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, 等. 基于模型的Atari强化学习。在*国际学习代表会议上*，2020年。

Anssi Kanervisto, Stephanie Milani, Karolis Ramanauskas, Nicholay Topin, Zichuan Lin, Junyou Li, Jianing Shi, Deheng Ye, Qiang Fu, Wei Yang, Weijun Hong, Zhongyue Huang、陈海成，曾广军，林跃，文森特-米切利，埃洛伊-阿隆索，弗朗索瓦-弗勒雷特，亚历山大-尼库林，尤里-贝卢索夫，奥列格-斯维琴科，和阿列克谢-什皮尔曼。Minerl钻石2021年竞赛：概述、结果和经验教训。In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, Proceedings of Machine Learning Research, 2022.URL <https://proceedings.mlr.press/v176/kanervisto22a.html>.

Steven Kapturowski, Georg Ostrovski, John Quan, Remi Munos, and Will Dabney.分布式强化学习中的循环经验回放。在*国际学习表征会议上*，2019年。

Andrej Karpathy. minGPT：对OpenAI GPT（生成性预训练转化器）训练的最小PyTorch再实现，2020年。URL <https://github.com/karpathy/minGPT>.

Levente Kocsis and Csaba Szepesvári. 基于强盗的蒙特卡洛规划. 在*欧洲机器学习会议上*, 2006年。

工藤琢和约翰-理查森. Sentencepiece: 一个用于神经文本处理的简单且独立于语言的子词标记器和去标记器。在*2018年自然语言处理中的经验方法会议上: 系统演示*, 第66-71页, 2018年。

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 使用学习到的相似度量进行超越像素的自动编码。在*国际机器学习会议上*, 第1558-1566页。PMLR, 2016.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: 强化学习的对比性无监督代表。在*国际机器学习会议上*, 第5639-5650页。PMLR, 2020.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel.应用于手写邮政编码识别的反向传播法. *神经计算*, 1 (4) : 541-551, 1989。

Christopher Manning和Anna Goldie.Cs224n自然语言处理与深度学习, 2022.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 通过深度强化学习实现人类水平控制. *自然》*, 518(7540):529-533, 2015.

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*国际机器学习会议上*, 第1928-1937页。PMLR, 2016.

Sherjil Ozair, Yazhe Li, Ali Razavi, Ioannis Antonoglou, Aaron Van Den Oord, and Oriol Vinyals. 用于规划的矢量量化模型. 在*国际机器学习会议上*, 第8302-8313页。PMLR, 2021.

Emilio Parisotto, Francis Song, Jack Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant Jayakumar, Max Jaderberg, Raphael Lopez Kaufman, Aidan Clark, Seb Noury, et al. Stabilizing transformers for reinforcement learning. 在*国际机器学习会议上*, 第7487-7498页。PMLR, 2020.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever.通过生成性预训练提高语言能力, 2018。

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever.语言模型是无监督的多任务学习者, 2019年。

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu.用一个统一的文本到文本的转化器探索转移学习的极限. *机器学习研究杂志*, 21:1-67, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever.0-shot文本-图像生成. 在*国际机器学习会议上*, 第8821-8831页。PMLR, 2021.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver.优先的经验回放. 在*国际学习表征会议*, 2016。

Martin Schmid, Matej Moravcik, Neil Burch, Rudolf Kadlec, Josh Davidson, Kevin Waugh, Nolan Bard, Finbarr Timbers, Marc Lanctot, Zach Holland, et al. *arXiv preprint arXiv:2112.03178*, 2021。

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, L. Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy P. Lillicrap, and David Silver.通过学习模型的计划掌握阿塔里、围棋、国际象棋和日本象棋. *自然》*, 588(7839): 604-609, 2020.

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。

迈克-舒斯特和中岛凯介。日语和韩语语音搜索.In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp.IEEE, 2012.

Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, and Philip Bachman.具有自我预测表征的数据高效强化学习。在*国际学习表征会议上*，2021年。

Younggyo Seo, Kimin Lee, Stephen L James, and Pieter Abbeel.通过视频进行无动作预训练的强化学习。在*国际机器学习会议上*，第19561-19579页。PMLR, 2022.

- David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *自然*, 529 (7587): 484-489, 2016。
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self play. *Science*, 362(6419): 1140-1144, 2018.
- Richard S. Sutton 和 Andrew G. Barto. *Reinforcement Learning : An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018.
- Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 注意力是你所需要的一切。 *神经信息处理系统的进展*, 30, 2017。
- Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *自然*, 575(7782):350-354, 2019.
- Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. 深度强化学习的决斗网络架构。在 *机器学习国际会议上*, 第1995-2003页。PMLR, 2016.
- 吴晨飞, 黄伦, 张倩茜, 李斌阳, 纪磊, 杨帆, Guillermo Sapiro, 段楠。Godiva: 从自然描述中生成开放域视频. *arXiv预印本arXiv:2104.14806*, 2021.
- Roman V Yampolskiy. *Artificial Intelligence Safety and Security*. Chapman & Hall/CRC, 2018.
- Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: 使用vq-vae和变换器的视频生成。 *arXiv预印本arXiv:2104.10157*, 2021。
- Denis Yarats, Ilya Kostrikov, and Rob Fergus. 图像增强是你所需要的一切: 从像素中规范化深度强化学习。在 *国际学习表征会议上*, 2021年。
- Weirui Ye, Shaohuai Liu, Thanard Kurutach, Pieter Abbeel, and Yang Gao. 用有限的数据掌握Atari游戏. *神经信息处理系统的进展*, 34, 2021。
- Qinqing Zheng, Amy Zhang, and Aditya Grover. 在线决策转化器.在 *国际机器学习会议上*, 第27042-27059页。PMLR, 2022.

A 模型和超参数

A.1 离散自动编码器

我们的离散自动编码器是基于VQGAN (Esser等人, 2021) 的实现。我们去掉了判别器, 基本上把VQGAN变成了一个带有额外感知损失的VQVAE (Van Den Oord等人, 2017) (Johnson等人, 2016; Larsen等人, 2016)。

培训目标如下:

$$\mathcal{L}(E, D, E) = \|x - D(z)\|_1 + \text{sg}(E(x) - E(z)) + \frac{1}{2} \text{sg}(E(z) - E(x)) + \mathcal{L}_{\text{感知性}}(x, D(z))$$

这里, 第一项是重建损失, 接下来的两项构成承诺损失 (其中sg(-)是停止梯度算子), 最后一项是知觉损失。

表2: 编码器/解码器的超参数。我们列出了编码器的超参数, 同样的参数也适用于解码器。

超参数	价值
框架尺寸 (高, 宽)	64 × 64
层数	4
每层的剩余区块	2
漩涡中的通道	64
自我注意层的分辨率	8 / 16

表3: 嵌入表的超参数。

超参数	价值
词汇量(N)	512
每帧代币(K)	16
符号嵌入维度 (d)	512

请注意, 在真实环境中收集经验时, 帧仍然要经过自动编码器以保持策略的输入分布不变。详见算法1。

A.2 变频器

我们的自回归变换器是基于minGPT (Karpathy, 2020) 的实现。它将 $L(K+1)$ 标记的序列作为输入, 并使用 $A \times D$ 嵌入表将其嵌入到 $L(K+1) \times D$ 张量中, 用于行动, 以及 $N \times D$ 嵌入表用于框架标记。这个张量通过 M 个转化器块被转发。我们使用类似GPT2的块 (Radford等人, 2019), 即

每个区块由一个自我注意模块组成, 该模块对输入进行了层级归一化处理, 用一个残差连接包裹, 然后是一个每个位置的多层感知器, 对输入进行层级归一化处理, 用另一个残差连接包裹。

表4: 转化器超参数

超参数	价值
时间步数 (L)	20
嵌入尺寸(D)	256
层数 (M)	10
负责人注意	4
重量衰减	0.01
嵌入辍学	0.1
注意力下降	0.1
剩余的辍学者	0.1

A.3 演员-CRITIC

除了最后一层，演员和评论家的权重是共享的。演员-评论家将 $64 \times 64 \times 3$ 的帧作为输入，并通过卷积块和LSTM单元进行转发（Mnih等人，2016；Hochreiter & Schmidhuber，1997；Gers等人，2000）。卷积块由以下部分组成

同一层重复四次： 3×3 卷积，跨度1和填充1，ReLU激活， 2×2 最大池，跨度2。LSTM隐藏状态的维度为512。在从一个给定的帧开始想象程序之前，我们对之前的20个帧进行烧录（Kapturowski等人，2019）以初始化隐藏状态。

表5: 训练循环和共享的超参数

超参数	价值
纪元	600
# 采集历时	500
每一纪元的环境步骤	200
收集epsilon-greedy	0.01
评估采样温度	0.5
在 epochs 之后启动自动编码器	5
纪元后启动变压器	25
纪元后启动演员批评法	50
自动编码器批量大小	256
变压器批量大小	64
演员-批评家的批量大小	64
每个周期的训练步骤	200
学习率	$1e-4$
优化器	亚当
亚当 β_1	0.9
亚当 β_2	0.999
最大梯度规范	10.0

B 行为者-批评者的学习目标

我们遵循Dreamer (Hafner等人, 2020; 2021) 的做法, 使用平衡偏差和方差的通用 λ -回报作为价值网络的回归目标。给定一个想象的轨迹 $(\mathbf{x}_0^{\wedge}, a_0, r_0^{\wedge}, d_0^{\wedge}, \dots, \mathbf{x}_{H-1}^{\wedge}, a_{H-1}, r_{H-1}^{\wedge}, d_{H-1}^{\wedge}, \mathbf{x}_H^{\wedge})$, λ -回报可以递归地定义为:
低点:

$$\Lambda_t = \begin{cases} r_t^{\wedge} + \gamma(1-d_t^{\wedge})(1-\lambda)V(\hat{\mathbf{x}}_{t+1}) + \lambda\Lambda_{t+1} & \text{如果 } t < H \\ V(\mathbf{x}_H^{\wedge}) & \text{如果 } t = H \end{cases} \quad (4)$$

价值网络 V 的训练是为了最小化 L_V , 即在想象的轨迹上与 λ -回报的预期平方差。

$$L_V = \mathbb{E}_{\pi} \sum_{t=0}^{H-1} (V(\mathbf{x}_t^{\wedge}) - \text{sg}(\Lambda_t))^2 \quad (5)$$

这里, $\text{sg}(-)$ 表示梯度停止操作, 意味着目标是基于梯度的优化中的一个常数, 这在文献中是经典的 (Mnih等人, 2015; Hessel等人, 2018; Hafner等人, 2020)。

由于在想象的MDP中产生了大量的轨迹, 我们可以为政策使用一个直接的强化学习目标, 如REINFORCE (Sutton & Barto, 2018)。为了减少REINFORCE梯度的方差, 我们使用值 $V(\mathbf{x}_t^{\wedge})$ 作为基线 (Sutton & Barto, 2018)。我们还增加了一个加权熵最大化的目标, 以保持充分的探索。行为者被训练成在想象的轨迹上最小化以下REINFORCE目标:

$$L_{\pi} = -\mathbb{E}_{\pi} \sum_{t=0}^{H-1} \log(\pi(a_t | \mathbf{x}_{\leq t}^{\wedge})) \text{sg}(\Lambda_t - V(\mathbf{x}_t^{\wedge})) + \eta H(\pi(a_t | \mathbf{x}_{\leq t}^{\wedge})) \quad (6)$$

表6: RL训练超参数

超参数	价值
想象地平线 (H)	20
γ	0.995
λ	0.95
η	0.001

C 可行性差距

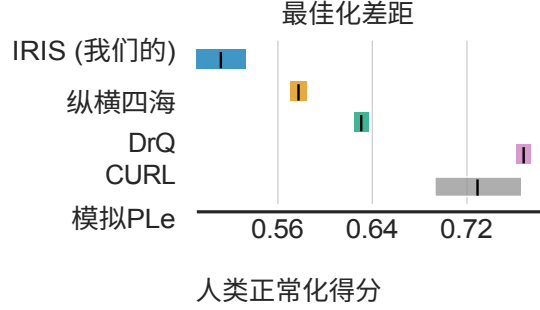


图8: 优化差距, 越低越好。算法未能达到人类水平得分的数量 (Agarwal等人, 2021)。

D IRIS ALGORITHM

算法1: IRIS

```

程序 training_loop(): 对于
    epochs 做
        collect_experience(step_collect)
        for steps_world_model do
            update_world_model()

            对于 step_behavior 做
                update_behavior()

程序 collect_experience(n):
    x0 ← env.reset()
    对于 t = 0 到 n - 1 做
         $\hat{x}_t \leftarrow D(E(x_t))$  // 通过离散自动编码器转发框架
        抽取一个  $\hat{a}_t \sim \pi(a_t | \hat{x}_t)$ 
        xt+1, rt, dt ← env.step(at)
        如果 dt = 1, 那么
            xt+1 ← env.reset()
    D ← D ∪ {xt, at, rt, dt}t=0n-1

程序 update_world_model():
    样本 {xt, at, rt, dt}t=τLτ+L-1 ~ D
    计算 zt := E(xt) 和  $\hat{x}_t^c := D(z_t)$  for t = τ, ..., τ + L - 1
    更新 E 和 D
    计算 pG(zt+1, rt, dt | zτ, aτ, ..., zt, at) t = τ, ..., τ + L - 1
    更新 G

程序 update_behavior():
    样本 x0 ~ D
    z0 ← E(x0)
     $\hat{x}_0^c \leftarrow D(z_0)$ 
    对于 t = 0 到 H - 1 做
        抽取一个  $\hat{a}_t \sim \pi(a_t | \hat{x}_t^c)$ 
        样本 zt+1, rt, dt ~ pG(zt+1, rt, dt | z0, a0, ..., zt, at)
         $\hat{x}_{t+1}^c \leftarrow D(z_{t+1}^c)$ 
    计算 V( $\hat{x}_t$ ) 为 t = 0, ..., H
    更新 π 和 V

```

E 具有不同数量标记的自动应答框架

转化器的序列长度由用于编码单帧的标记数和内存中的时间步数决定。增加每一帧的标记数会带来更好的重建效果，尽管它需要更多的计算和内存。

这种权衡在具有大量可能配置的视觉挑战游戏中尤为重要，在这些游戏中，离散自动编码器很难对只有16个标记的帧进行正确编码。例如，图9显示，当在《异形》中把每一帧的代币数量增加到64个时，离散自动编码器正确地重建了玩家、其敌人和奖励。

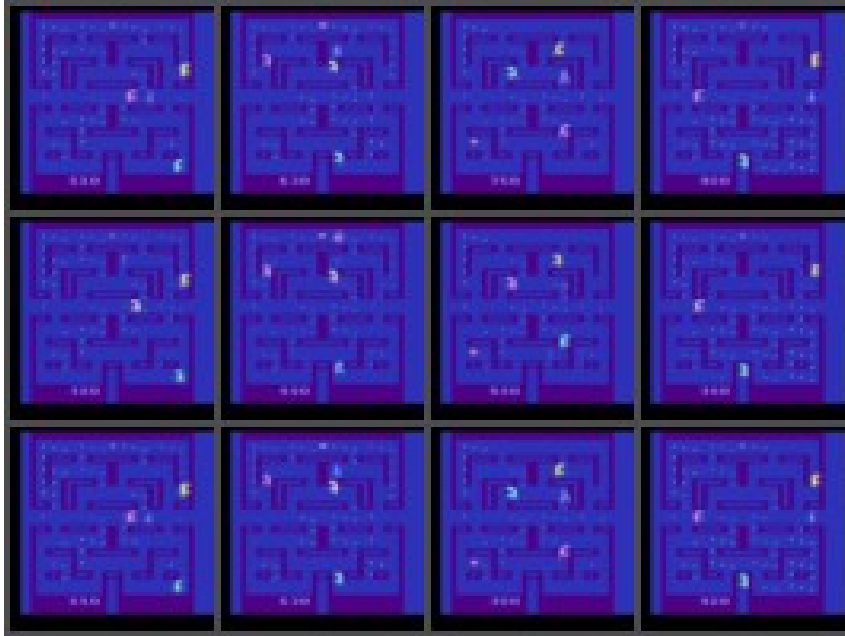


图9：在《异形》中，每帧标记的数量和重建的质量之间的权衡。每一列显示的是真实环境中的64×64帧（顶部），用16个标记的离散编码进行的重建（中间），以及用64个标记的离散编码进行的重建（底部）。

在《异形》中，玩家是深蓝色的角色，而敌人是大型彩色精灵。在每帧16个代币的情况下，自动编码器经常会擦掉玩家，切换颜色，并把奖励放错位置。当增加代币的数量时，它可以正确地重建帧。

表7显示了IRIS在三款游戏中以每帧64个代币训练的最终表现。有趣的是，尽管世界模型更准确，但在《异形》中的表现只略有增加（+36%）。这一观察表明，《异形》构成了一个困难的强化学习问题，其他基线在该游戏中的低表现就证明了这一点。相反，IRIS在Asterix（+121%）和BankHeist（+432%）中因每帧有更多的代币而大大受益。

表7：每帧64个代币而不是16个代币的Alien、Asterix和BankHeist的回报。

游戏	随机的人类SimPLe CURL				DrQ	SPR IRIS (16枚硬币)	IRIS (64枚硬币)
外星人	227.8	7127.7	616.9	711.0	865.2	841.9	420.0
阿斯特里克斯	210.0	8503.3	1128.3	567.2	763.6	962.5	853.6
银行劫案	14.2	753.1	34.2	65.3	232.9	345.4	53.1
							282.5

F 超越样本效率的设定

IRIS可以通过增加用于编码帧的标记数量、增加模型的容量、每个环境步骤采取更多的优化步骤或使用更多的数据来扩大规模。在这个实验中，我们通过将环境步骤的数量从100k增加到10M来研究数据的扩展性。然而，为了在我们的计算资源范围内保持训练时间，我们将每个环境步骤的优化步骤比例从1:1降低到1:50。因此，在10万帧的情况下，这个实验的结果会比论文中报告的结果更差。

表8：将环境步骤的数量从100K增加到10M。

游戏	随机	人类	IRIS (100k)	IRIS (10M)
外星人	227.8	7127.7	420.0	1003.1
Amidar	5.8	1719.5	143.0	213.4
殴打	222.4	742.0	1524.4	9355.6
阿斯特里克斯	210.0	8503.3	853.6	6861.0
银行劫案	14.2	753.1	53.1	921.6
战斗地带	2360.0	37187.5	13074.0	34562.5
拳击	0.1	12.1	70.1	98.0
突围赛	1.7	30.5	83.7	493.9
斩波器命令	811.0	7387.8	1565.0	9814.0
疯抢者	10780.5	35829.4	59324.2	111068.8
恶魔的攻击	152.1	1971.0	2034.4	96218.6
高速公路	0.0	29.6	31.1	34.0
冻伤	65.2	4334.7	259.1	290.3
地鼠	257.6	2412.5	2236.1	97370.6
英雄	1027.0	30826.4	7037.4	19212.0
詹姆斯邦德	29.0	302.8	462.7	5534.4
袋鼠	52.0	3035.0	838.2	1793.8
克鲁尔	1598.0	2665.5	6616.4	7344.0
巩俐老师	258.5	22736.3	21759.8	39643.8
MsPacman	307.3	6951.6	999.1	1233.0
乒乓	-20.7	14.6	14.6	21.0
私密的眼睛	24.9	69571.3	100.0	100.0
Qbert	163.9	13455.0	745.7	4012.1
路人甲	11.5	7845.0	9614.6	30609.4
海底捞	68.4	42054.7	661.3	1815.0
向上向下	533.4	11693.2	3546.2	114690.1
#超级人类(个)	0	不适用	10	15
平均值(个)	0.000	1.000	1.046	7.488
中位数(个)	0.000	1.000	0.289	1.207
IQM (个)	0.000	1.000	0.501	2.239
最优化差距 (↓)	1.000	0.000	0.512	0.282

表8显示，将环境步骤的数量从100k增加到10M，极大地提高了大多数游戏的性能，提供了IRIS可以扩大到超过样本效率制度的证据。在一些游戏中，更多的数据只能产生微弱的改善，这很可能是由于困难的探索问题或视觉上的挑战领域，这将受益于更多的代币来编码帧（附录E）。

G 计算资源

对于每个Atari环境，我们用5个不同的随机种子反复训练IRIS。我们用8个Nvidia A100 40GB GPU进行实验。在同一个GPU上运行两个Atari环境，训练需要大约7天，结果每个环境平均需要3.5天。

SimPLe (Kaiser等人, 2020) 是唯一涉及想象力学习的基线，在单一环境下用P100 GPU训练了3周。至于SPR (Schwarzer等人, 2021)，最强的基线，没有前瞻搜索，它用P100的GPU在4.6小时内进行了明显的快速训练。

关于带有前瞻搜索的基线，MuZero (Schrittwieser等人, 2020) 最初使用40个TPU在单一的Atari环境中训练了12个小时。Ye等人 (2021) 用4个RTX 3090 GPU在7个小时内同时训练EfficientZero和他们对MuZero的重新实现。EfficientZero的实现依赖于CPU和GPU线程并行运行的分布式基础设施，以及MCTS的C++/Cython实现。相比之下，IRIS和没有超前搜索的基线依靠的是直接的单GPU/单CPU实现。

H 在高速公路上探索

Freeway中的奖励功能是稀疏的，因为代理人只有在完全越过公路时才会得到奖励。此外，撞到汽车会拖累它，使它不能顺利地登上公路。这给新初始化的代理带来了一个探索问题，因为随机策略几乎肯定不在10万帧的预算下获得非零的奖励。



图10：一局《高速公路》。汽车会把玩家撞倒，使其很难穿越马路并获得随机政策的奖励。

这个问题的解决方案其实很简单，只需要在UP动作被超采样时延长时间。大多数Atari 100k基线用epsilon-greedy时间表和argmax动作选择来解决这个问题，在某些时候，网络配置会使UP动作受到很大的青睐。在这项工作中，我们选择了更简单的策略，即有一个固定的epsilon-greedy参数并从策略中取样。然而，我们将采样温度从1降低到

0.01，以避免在训练的早期阶段出现不利于学习的随机漫步。因此，一旦它通过探索获得了最初的几个奖励，IRIS就能够在其世界模型中内化稀疏的奖励函数。