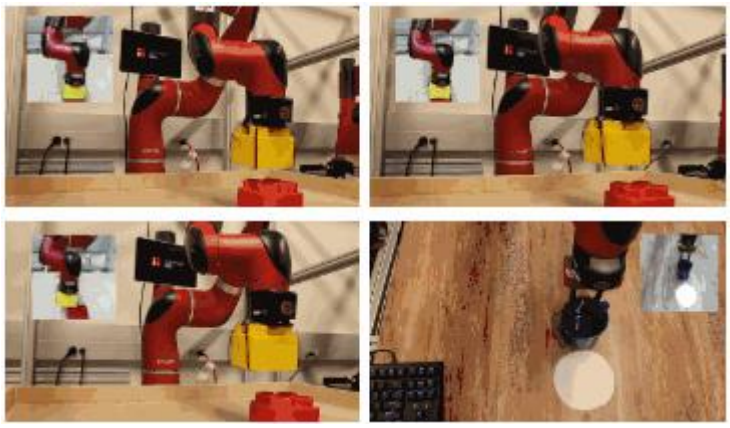
机器人基于图像完成任务最有效的 RL 方法：无需预测未来，也无需严格假设！

人工智能7 月 12 日

对于机器人强化学习来说，基于视觉的物块堆叠和推动是最常见的任务，为了减小训练过程的成本和安全问题，我们希望最小化训练过程中与环境交互的次数。但从相机这样复杂的图像传感器中进行高效学习却十分困难。为了解决这一问题，伯克利的研究人员提出了一种新型基于模型的强化学习方法并发表了相关文章介绍了这一成果，雷锋网 AI 科技评论将其编译如下。

概述

想象一下这样的场景：一个机器人试图通过相机影像的视觉输入来学习堆叠物块和推动物体。为了最大限度地降低成本和安全问题，我们希望能够最小化机器人的交互学习时间，但从相机这样复杂的图像传感器中进行高效学习依旧十分困难。因此本工作提出了 SOLAR——一种新的基于模型的增强学习（RL）方法，它直接从视觉输入和不到一小时的交互中学习技能，包括在真正的 Sawyer 机械臂上执行高难度任务。据我们所知，SOLAR 是解决机器人在现实世界中基于图像完成任务的最有效的 RL 方法。



机器人使用 SOLAR 一个小时内学会了如何堆积木和推杯子

在 RL 设置中，机器人通过反复试错从自己的经验中学习，以最大限度地降低与当前任务相对应的成本函数。近年来，许多具有挑战性的任务都是通过 RL 方法解决的，但这些成功案例大多来自无模型（model-free）的 RL 方法，与基于模型（model-based）的方法相比，这些方法通常需要更多的数据。然而，基于模型的方法往往依赖于精准预测未来的能力，以便规划主体的操作。对于基于图像学习的机器人来说，预测未来的图像本身需要大量的交互训练，因此我们需要解决这个问题。

一些基于模型的 RL 方法不需要精准的未来预测，但这些方法通常会对状态进行严格的假设。LQR-FLM（linear-quadratic regulator fitted linear models，https://arxiv.org/abs/1504.00702）方法已被证明可以通过对系统动力学状态的近似线性假设来高效地学习新的任务，这个方法同样可适用于大部分机器人系统。然而，这种假设对于基于图像的学习，却是令人望而却步的，因为相机反馈的像素动态远不是线性能够表达的。因此，我们在工作中研究的问题是如何放宽这一假设，以便开发得到一种基于模型的 RL 方法，在无需精准未来预测的情况下解决基于图像的机器人任务呢？

最后，我们通过使用深层神经网络学习潜在状态表示来解决这个问题。当机器人处理来自任务的图像时，它可以将图像编码为潜在表示，然后将其用作 LQR-FLM 的状态输入来代替图像本身。其中的关键在于 SOLAR 模型可以学习紧凑的潜在状态表示，从而实现对目标的精确捕捉；然后模型通过鼓励潜在状态的动力学倾向于线性表示，来学习到一种可以有效用于 LQR-FLM 算法的表示。为此，我们引入了一个明确表示潜在线性动力学的潜在变量模型，将该模型与 LQR-FLM 相结合，为 SOLAR 算法提供了基础。

潜在表示的随机最优控制

SOLAR（stochastic optimal control with latent representations）意为具有潜在表示的随机最优控制，它是基于图像 RL 设置的一种有效且通用的解决方案。SOLAR 的关键在于它可以学习线性动力学精准的潜在状态表示，并利用了不依赖于未来预测的基于模型的 RL 方法。

线性动态控制

控制理论中最著名的结果之一是线性二次型调节器（LQR），这是一组方程式，为线性动力学且二次型的系统提供最优控制策略。虽然现实世界的系统几乎从不是线性的，但是 LQR 的近似值，例如具有拟合线性模型（LQR-FLM）的 LQR 已被证明在各种机器人控制任务中表现良好。与其他基于模型的 RL 方法相比，LQR-FLM 一直是学习控制技能最有效的 RL 方法之一。线性模型的简单性以及这些模型不需要准确预测未来的特点使得 LQR-FLM 成为一种吸引人的构建方法，但是这种方法的关键限制是它通常假定访问系统状态，例如机器人的关节配置和感兴趣对象的位置，这

通常是合理地建模为近似线性。我们通过学习可以用作 LQR-FLM 输入的表示来替代图像并放宽这个假设。

使用系统状态，LQR-FLM 和相关方法已被用于成功学习无数的任务，包括机器人操纵和运动。我们的目标是通过自动学习从图像到 LQR-FLM 的状态输入来扩展这些功能。

从图像中学习潜在状态

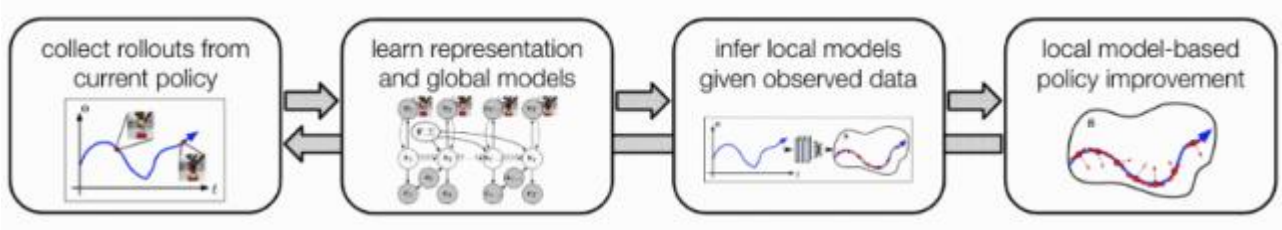
我们建立的图形模型假设我们观察到的图像是潜在状态的函数，并且状态根据由行为的线性动力学调制，损失由状态和行为的二次函数给出。

我们希望智能体可以从其视觉输入中提取一种状态表示，其中的状态动态尽可能接近线性。我们设计了一个潜在变量模型来实现，其中潜在状态服从线性动力学，如下图所示。深色节点是我们从与环境交互中观察到的图像、行为与成本。浅色节点代表系

统基本状态，这是我们希望学习的表示形式，我们假设下一个状态是由当前状态和操作的线性函数所得。该模型与结构化变分自编码器（structured variational auto-encoder）有很强的相似之处，该模型以前适用于表示老鼠视频的特征等应用。我们用来适应模型的方法也是基于前面工作中提出的方法。

在较高的层级上，该方法将同时学习状态动力学和编码器，将当前和先前图像作为输入来估计当前状态。如果我们对多个机器人与环境的交互相对应的观察图像序列进行编码，可以看到这些状态序列是否匹配学到的线性动力学行为；如果它们不这样做，我们将调整动力学和编码器，使它们估计所得状态向线性逼近。该过程的关键在于我们不是直接优化模型来使预测时更精准，而是调整线性模型匹配主体先前与环境的交互。这有力地弥补了 LQR-FLM 的不足，使得它也不依赖预测就能获得良好的性能。关于该模型学习流程的更多细节，请前往以下地址参考我们的论文：https://arxiv.org/abs/1808.09105 。

SOLAR 算法



我们的机器人迭代地与其环境交互，使用此数据更新其模型，使用此模型来估计潜在状态及其动态，并使用这些动态更新其行为。

现在我们已经阐述了该方法的创建步骤，那这些步骤如何一同组合成 SOLAR 方法呢？智能体根据策略在环境中运作，而策略则根据当前的潜在状态估计来指导行动。

这些交互产生的图像、行为和损失的轨迹，再用于匹配动力学模型。之后，利用这些完整的交互轨迹，我们的模型不断完善它对潜在动态的估计，这使得 LQR-FLM 能够产生一个更新的策略，该策略将使得智能体在给定的任务中表现得更好，即降低损失（成本）。然后更新的策略将用于收集更多交互轨迹以及不断重复优化过程。上图展示该算法的各个阶段。

LQR-FLM 和大多数其他基于模型的 RL 方法相比，关键区别在于生成的模型仅用于策略改进，而不用于预测未来。这在观测复杂且难以预测的场景中非常有用，我们通过引入可与动力学一同估计的潜在状态，将这一优势扩展到基于图像的场景中。最终，SOLAR 只需使用一个小时与环境交互，即可为基于图像的机器人操作任务生成良好的策略。

实验

我们主要在 Sawyer 机械臂上测试了 SOLAR，其中机械臂有七度的自由度，可用于各种操作任务。我们给机械臂输入来自对准其手臂的摄像头的图像以及场景中的相关物体，然后让机械臂完成学习乐高方块堆叠和推动杯子的任务。

左：对于乐高积木堆叠，我们试验了臂和块的多个起始位置。（对于推动，我们只使用人类在机器人成功时按下键提供的稀疏奖励，示例图像观察在底行给出。）右：SOLAR 学习的成功行为示例。

乐高方块堆叠任务

块堆叠的主要挑战来自于成功完成任务所需的精度，因为机械臂必须非常准确地放置块，才能将各个模块衔接起来。在 SOLAR 系统下，Sawyer 只需从输入的相机镜头中学习这种精度，与此同时它还能成功掌握从手臂和积木的多个起始积木位置中学习堆叠。

其中，当积木的起始位置在桌面上，是最具挑战性的，因为 Sawyer 必须先将积木从桌子上拿起，然后再堆叠它，即它无法变得「贪婪」，更无法简单地径直将积木移向另外的积木。

我们首先将 SOLAR 当作使用标准变分自编码器（VAE）而非结构化变分自编码器（SVAE）的消融方法，这意味着学习到的状态表示不再遵循线性动力学。而这种消融的方法，机械臂只有在最简单的起始位置的前提下才能取得成功。为了理解模型无需精准预测未来给我们所带来的益处，我们将 SOLAR 比作另一种消融方法，即使用一种可供替代的规划方法——模型预测控制模型（MPC）来替代 LQR-FLM 算法，同时我们也将其视作此前使用了 MPC 的一种性能最佳的方法，即深度视觉预见（DVF，https://arxiv.org/abs/1812.00568 ）。其中，MPC 常被应用于此前和随后的一系列工作中，并且它依赖于使用学习到的模型来生成精确的未来预测的能力，从而决定需要采取什么样的行动来提升性能。

虽然 MPC 消融在两个更简单的初始位置上学习得更快，但它无法应对最为复杂的场景，因为 MPC 只能「贪婪地」缩短两个积木之间的距离，而无法将积木从桌面上拿下来。MPC 之所以贪婪地行动，是因为它仅能进行短期规划，而长远来看，它

对未来图像的预测则会越来越不精准，这恰恰就是 SOLAR 能够利用 LQR-FLM 来完全避免进行未来预测从而克服的失败的方式。之后，我们发现 DVF 虽然取得了一定的进步，但是最终并不能解决这两个更加困难的场景，即便在比我们方法使用更多数据的情况下。这证明了我们的方法具有更高的数据效率，可以在几个小时内实现 DVF 需要几天甚至几周才能解决的问题。

杯子推动任务

此外我们还研究了机械臂在推动杯子任务上的表现。我们通过用稀疏的奖励信号替换成本来增加机械臂推动杯子时的额外挑战，比如说机械臂只有完成了任务时才会被告知信号，否则就不会被告知。如下图所示，人类在键盘上按下一个键来提供稀疏的奖励信号，而机械臂需要推理出如何改进行为来获得这一奖励。我们通过对 SOLAR 进行直接拓展便处理了这一问题，详细内容可参考我们的论文（论文地址：https://arxiv.org/abs/1808.09105 ）。即便面临着额外的挑战，我们的方法在一个小时左右的交互后也成功地学习到了推动杯子的策略，大大超过了相同数据量下 DVF 的表现。

模拟比较

除了 Sawyer 实验之外，我们还在模拟中进行了几次比较，因为大多数先前的工作并未尝试使用真正的机器人进行实验。特别地，我们建立了一个 2D 导航域，其中底层系统实际上具有线性动力学和二次成本，但我们只能观察显示智能体和目标的自上而下视图的图像。我们还包括两个更复杂的域：一辆必须从 2D 平面右下角驱动

到左上角的汽车，以及一个负责达到左下角目标的 2 自由度机械臂。所有域都是通过只提供任务自上而下视图的图像观察来学习的。

我们比较了鲁棒局部线性可控嵌入（RCE，https://arxiv.org/abs/1710.05373 ），它采用不同的方法来学习遵循线性动力学的潜在状态表示。我们还将其与近端策略优化（PPO）进行了比较，PPO 是一种无模型 RL 方法，用于解决许多模拟机器人领域问题，但这种方法对于现实世界学习而言，数据效率不够高。我们发现 SOLAR 比 RCE 学习速度更快，最终性能更好。PPO 通常能比 SOLAR 学习到更好的最终性能，但这通常需要 1 到 3 个数量级的数据，这对于大多数现实机器人的学习任务来说也是可望不可及的。这种权衡是普遍存在的：无模型方法往往会获得更好的最终性能，但基于模型的方法学得更快。

相关工作

学习图像潜在表示的方法提出了类如重建图像和预测未来图像等的目标。这些目标并不完全符合我们完成任务的目标，例如机器人在按颜色将目标分类到垃圾箱中时，并不需要完美地重建他前面的墙壁的颜色。我们还开展了适合于控制的状态表示方面的工作，包括识别图像中的兴趣点和学习潜在状态，从而使各个维度独立控制。最近的一篇调查论文还对状态表示学习的前景进行了分类。

除了控制之外，我们最近还进行了大量学习数据结构化表示的工作，其中许多工作扩展了 VAE。SVAE 就是一个这种框架的例子，其他一些方法也试图用线性动力学来解释数据。除此之外，还有一些研究通过混合模型结构、各类离散结构和贝叶斯非参数结构来学习潜在表示。

我们还提出了与我们在之前和随后的工作中提出的观点密切相关的想法。如前所述，DVF 还直接从视觉中学到了机器人任务，最近的一篇博客文章（文章查看地址：https://bair.berkeley.edu/blog/2018/11/30/visual-rl/ ）总结了该结果。嵌入控制及其后继的 RCE 还旨在学习线性动力学的潜在状态表示。我们在论文中将这些方法进行了比较，并证明了我们的方法往往表现出更好的性能。在我们的成果之后，研究人员提出的 PlaNet 混合利用确定性和随机变量来学习潜在状态表示 ,并将它们与 MPC 结合使用，其中，MPC 是我们评估中的基准方法之一，在几个模拟任务上展示了良好的结果。正如实验所显示，LQR-FLM 和 MPC 各有优缺点，我们发现 LQR-FLM 通常在机器人控制方面更为成功，避免了 MPC 的贪婪行为。

未来的工作

我们看到了未来工作的几个令人兴奋的方向，在此简要提及两个方向：

首先，我们希望我们的机器人能够学习复杂、多阶段的任务，例如构建乐高结构而不仅仅是堆叠一个个方块，或进行更复杂的推动任务而不仅仅是推动一个杯子。我们可以通过提供所希望机器人完成目标的中间图像来实现这一点，如果我们期望机器人能够分别学习每个阶段，这一算法也许能够将这些策略串在一起，形成更复杂、更有趣的行为。

其次，人类不仅学习状态的表示，而且还学习动作——我们不考虑单个肌肉运动，而是将这些运动组合成「宏观动作」，以执行高度协调和复杂的行为。如果我们能够类似地学习动作表示 ,我们就能使机器人更有效地学习如何使用硬件 ,比如说灵巧的手，这将进一步提高他们处理复杂的现实环境的能力。

原文链接

https://bair.berkeley.edu/blog/2019/05/20/solar/

- END -