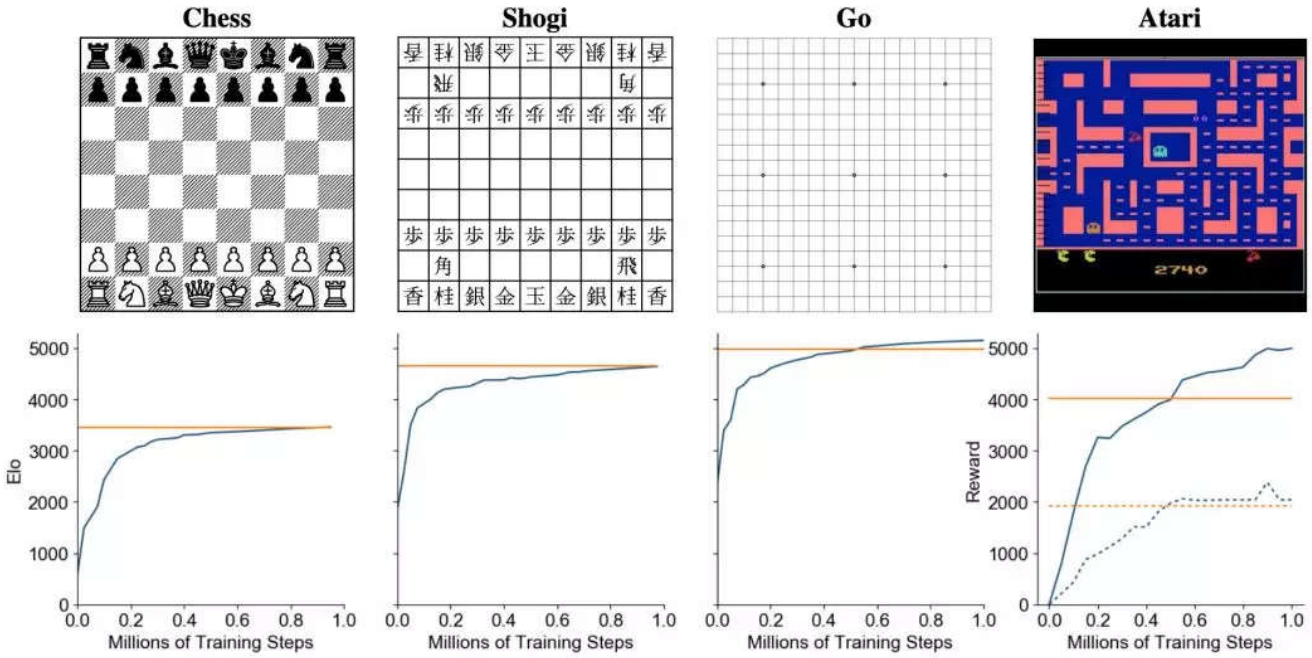
通用AlphaGo诞生？DeepMind的MuZero在多种棋类游戏中超越人类

机器学习研究会订阅号 前天

> DeepMind近期的一项研究提出了MuZero算法，该算法在不具备任何底层动态知识的情况下，通过结合基于树的搜索和学得模型，**在雅达利2600游戏中达到了SOTA表现，在国际象棋、日本将棋和围棋的精确规划任务中可以匹敌AlphaZero，甚至超过了提前得知规则的围棋版AlphaZero。**



MuZero 算法在国际象棋、日本将棋、围棋和雅达利（Atari）游戏训练中的评估结果。横坐标表示训练步骤数量，纵坐标表示 Elo评分。黄色线代表 AlphaZero（在雅达利游戏中代表人类表现），蓝色线代表 MuZero。

基于前向搜索的规划算法已经在 AI 领域取得了很大的成功。在围棋、国际象棋、西洋跳棋、扑克等游戏中，人类世界冠军一次次被算法打败。此外，规划算法也已经在物流、化学合成等诸多现实世界领域中产生影响。然而，这些规划算法都依赖于环境的动态变化，如游戏规则或精确的模拟器，导致它们在机器人学、工业控制、智能助理等领域中的应用受到限制。

基于模型的强化学习旨在通过以下步骤解决这一问题：首先学习一个环境动态模型，然后根据所学模型进行规划。一般来说，这些模型要么着眼于重建真实的环境状态，要么着眼于完整观察结果的序列。然而，之前的研究在视觉上丰富的领域还远远没有达到 SOTA 水准，如雅达利 2600 游戏。

最受欢迎的方法是基于无模型强化学习的方法，即直接从智能体与环境的交互中估计优化策略和/或价值函数。但在那些需要精确和复杂前向搜索的领域（如围棋、国际象棋），这种无模型的算法要

远远落后于 SOTA。

研究者在57个不同的雅达利游戏中评估了MuZero，发现该模型在雅达利2600游戏中达到了SOTA表现。此外，他们还在不给出游戏规则的情况下，在国际象棋、日本将棋和围棋中对MuZero模型进行了评估，发现该模型可以匹敌AlphaZero超越人类的表现。而且，在该实验中，其前辈AlphaZero提前获知了规则。

MuZero 算法概览

MuZero 基于 AlphaZero 强大的搜索和基于搜索的策略迭代算法，但又将一个学习好的模型整合到了训练步骤中。MuZero 还将 AlphaZero 扩展到了一个更加广泛的环境集合，包含单个智能体域和中间时间步上的非零奖励。

该算法的主要思路是预测那些与规划直接相关的未来行为（如下图 1 所示）。模型将接收到的观察结果（如围棋棋盘图像或雅达利游戏截图）作为输入，然后将其转换为一个隐藏状态。接下来，通过一个循环过程来迭代更新该隐藏状态，该循环过程接收前一个隐藏状态和假设的下一步操作。

在每一个步骤上，模型会预测策略（如玩的动作）、价值函数（如预测的赢家）以及即时奖励。对模型进行端到端训练的唯一目标是准确估计这三个重要的量，以匹配改进的策略估计和通过搜索及观察到的奖励生成的值。

对于隐藏的状态，没有直接的约束和要求来捕获重建原始观察结果所需的信息，大大减少了模型维护和预测的信息量；也没有要求隐藏状态匹配环境中未知、真实的状态；更没有针对状态语义的其他约束。

相反，隐藏状态能够地以任何与预测当前和未来值和策略相关的方式来表示状态。直观地说，智能体可以在内部创建规则和动态，以实现最精确的规划。

图 1：用一个训练好的模型进行规划、行动和训练。（A）MuZero 利用其模型进行规划的方式；（B）MuZero 在环境中发生作用的方式；（C）MuZero 训练其模型的方式。

MuZero 算法详解

研究者对 MuZero 算法进行了更详细的解读。在每个时间步 t 上、以过往观察结果 O_1, ..., O_t 和未来行为 a_t+1, ..., a_t+k 为条件、通过一个具有参数θ的模型μ_θ，为每个 k=1...K 步进行预测。该模型预测三种未来数量：策略 、价值函数 和即时奖励 ，其中 u. 表示观察到的正确奖励，π表示用来选择实时行动的策略，γ表示环境的贴现函数（discount function）。

在每个时间步 t 上，MuZero 模型由表征函数、动态函数和预测函数联合表征。在本文中，研究者对动态函数进行了确切的表征。策略和价值函数则通过预测函数 并根据内部状态 s^k 来计算，这与 AlphaZero 的联合策略和价值网络相似。

给定这样一个模型，则有可能在基于过往观察结果 O_1, ..., O_t 的情况下查找基于假设的未来轨迹 a^1, ..., a^k。例如，一个简单的搜索可以轻松地选择最大化价值函数的 k 步动作序列。更普遍地说，我们或许可以将任何 MDP（马尔科夫决策过程）规划算法应用于由动态函数推导出的内部奖励和状态空间。

对于每个假设的时间步 k，模型的所有参数接受联合训练，从而在 k 个实际的时间步后，对策略、价值和奖励与它们各自对应的目标值进行精确的匹配。与 AlphaZero 相似，提升后的策略目标通过蒙特卡洛树（MCTS）搜索生成。第一个目标是最小化预测策略 p^k_t 和搜索策略π_t+k 之间的

误差；第二个目标是最小化预测值 v^k_t 和价值目标 z_t+k 之间的误差；第三个目标是最小化预测奖励 r^k_t 和观察到的奖励 u_t+k 之间的误差。最后添加 L2 正则化项，得出以下总损失：

实验结果

在实验中，研究者将 MuZero 算法应用于围棋、国际象棋和日本将棋等经典棋盘游戏中，作为挑战规划问题的基准；同时又应用于雅达利游戏环境中的 57 个游戏，作为视觉复杂强化学习领域的基准。

下图 2 展示了 MuZero 算法在每个游戏训练中的性能。在围棋游戏中，尽管搜索树中每个节点的计算量小于 AlphaZero，但 MuZero 的性能依然略微超过 AlphaZero。这表明 MuZero 可能在搜索树中缓存自身计算，并利用动态模型的每个附加应用来对位置产生更深的理解。

图 2：MuZero 算法分别在国际象棋、日本将棋、围棋和雅达利游戏训练中的评估结果。在国际象棋、日本将棋和围棋游戏中，横坐标表示训练步骤数量，纵坐标表示 Elo 评分。

表 1：雅达利游戏中 MuZero 与先前智能体的对比。研究者分别展示了大规模（表上部分）和小规模（表下部分）数据设置下 MuZero 与其他智能体的对比结果，表明 MuZero 在平均分、得分中位数、Env. Frames、训练时间和训练步骤五项评估指标**（红框）**取得了新的 SOTA 结果。

为了了解 MuZero 中模型的作用，研究者还重点在围棋和吃豆人雅达利游戏中进行了以下几项实验。他们首先在围棋的典型规划问题上测试了规划的可扩展性（下图 3A）。此外，他们还研究了所有雅达利游戏中规划的可扩展性（下图 3B）。接着，他们将自己基于模型的学习算法与其他相似的无模型学习算法进行了比较（下图 3C）。

图 3：MuZero 在围棋、57 个雅达利游戏、吃豆人游戏上的评估结果。

论文链接：https://arxiv.org/pdf/1911.08265.pdf

转自：机器之心