**Sim-to-real**

简介：

强化学习是一种根据环境反馈进行学习的技术。强化学习 agent 辨别自身所处的状态 (state)，按照某种策略决定动作（action），并根据环境提供的奖赏来调整策略，直至达到最优。

马尔可夫决策 MDP（Markov Decision Process）是强化学习任务的标准描述，我们定义一个任务 M，用四元组< S , A , T, R>表示，其中 S 是状态空间，A 是动作空间，T 是状态转移概率，R 是奖赏函数。state-action 空间 S×A 定义了任务的域，状态转移概率 T 和奖赏函数 R 定义了任务的目标。当强化学习的状态动作空间 S×A 很大时，为了寻找最优策略，搜索过程非常耗时。此外，学习近似最优解所需的样本数量在实际问题中往往令人望而却步。无论是基于值的方法还是基于策略的方法，只要问题稍稍变动，之前的学习结果就会失效，而重新训练的代价巨大。因此，研究者们针对强化学习中的迁移学习展开了研究，希望能够将知识从源任务迁移到目标任务以改善性能。

样本数量、学习结果针对不同任务失效等这些在一般强化学习中存在的问题在机器人强化学习中尤其突出。现有的机器人强化学习方法大多只能完成单个任务，而无法在不同的任务之间推广，或者仅通过收集很少数量的现实机器人经验来概括一些任务策略，其性能无法满足实战的要求。将迁移学习引入机器人强化学习，目的是利用模拟环境中的数据辅助现实机器人学习。机器人强化学习中的迁移学习称为「模拟到现实（Sim-to-real）方法」，具体是指首先在模拟环境中收集数据并训练机器人控制策略，然后进行迁移学习，将训练获得的控制策略（新技能）应用于物理现实中的机器人。

模拟器与物理世界之间的现实差距经常导致在与物理机器人一起工作时失败。这种差距是由物理参数（例如摩擦、kp、阻尼、质量、密度）之间的不一致性以及更严重的是错误的物理建模（例如软表面之间的碰撞）触发的。

为了弥补模拟与现实之间的差距，我们需要改进模拟器并使其更接近现实。有几种方法：

**•**[**系统辨识**](https://zhida.zhihu.com/search?content_id=232805485&content_type=Article&match_order=1&q=%E7%B3%BB%E7%BB%9F%E8%BE%A8%E8%AF%86&zhida_source=entity)**(System identification)**  
– 系统辨识是建立物理系统的数学模型的过程；在强化学习的背景下，数学模型即为模拟器。为了使模拟器更加逼真，需要进行仔细的校准。  
– 不幸的是，校准是昂贵的。此外，同一机器的许多物理参数可能会因温度、湿度、定位或时间的磨损而发生显著变化。

**•**[**领域自适应**](https://zhida.zhihu.com/search?content_id=232805485&content_type=Article&match_order=1&q=%E9%A2%86%E5%9F%9F%E8%87%AA%E9%80%82%E5%BA%94&zhida_source=entity)**(Domain adaptation)**  
– 领域自适应（DA）是指一组转移学习技术，通过任务模型施加的映射或正则化来更新模拟中的数据分布，使其与真实数据分布相匹配。  
– 许多 DA 模型，特别是用于图像分类或基于图像的端到端强化学习任务的模型，是基于对抗损失或GAN构建的。  
  
**• 领域随机化 (Domain randomization)**

– 通过领域随机化（DR），我们能够创建各种具有随机属性的模拟环境，并训练在所有这些环境中都有效的模型。  
– 可能该模型可以适应真实世界环境，因为真实系统预期是训练变化丰富性分布中的一个样本。  
DA 和 DR 都是无监督的方法。

相关发展：

**缩小模拟和现实的差距方法：**

在论文《Florian Golemo, Adrien Ali Taiga, et al, 「Sim-to-Real Transfer with Neural-Augmented Robot Simulation,」PMLR 2018,》提出的 sim-to-real（从仿真到现实）的方法是神经增强模拟（Neural-Augmented Simulation，NAS），旨在通过训练循环神经网络来缩小模拟与现实之间的差距，使在仿真环境中学习的控制策略能更好地迁移到现实环境中。

方法概述：

将源域（通常是模拟器）和目标域（现实世界中的机器人）看作不同的马尔可夫决策过程（MDP）。利用从现实机器人收集的数据训练长短期记忆模型（LSTM），学习源域和目标域轨迹之间的差异映射，进而校正模拟轨迹，使其更接近现实情况。

方法介绍：

目标域数据收集：依据行为策略（可以是随机策略或专家提供的策略），从目标环境（现实机器人）收集轨迹数据。在收集过程中，机器人根据执行动作，并记录相应的状态变化 。

训练策略：

从目标域分布中采样初始状态，让源域和目标域从相同初始状态开始。在每个时间步，依据行为策略采样动作，分别在源域和目标域执行该动作，获取状态转移信息。将这些数据用于训练 LSTM 模型，使其学习预测目标域状态与源域状态的差异，即输出校正项 。

迁移到目标域：

训练好模型后，将其与源环境结合学习策略。在学习过程中，将源域的当前状态转移信息输入，计算目标环境当前状态的估计值。根据该估计值选择动作，然后将源域状态设置为目标域状态的估计值，以此校正源域轨迹，使其更接近目标域轨迹

**随机化处理策略训练方法：**

在论文《Sim-to-Real Transfer of Robotic Control with Dynamics Randomization》中：

提出通过动力学随机化（Dynamics Randomization）实现 sim-to-real（从仿真到现实）的迁移，以解决强化学习在真实物理系统应用中的难题。

核心思路：利用低精度仿真环境，在训练过程中对模拟器的动力学参数进行随机化处理，使智能体学习到能适应多种动力学情况的策略，从而增强策略对现实世界动力学的泛化能力，实现从仿真到现实的直接迁移。

**方法介绍：**

**动力学随机化：**

在训练时，每次 episode 开始前从指定分布中随机采样一组动力学参数，这些参数涵盖机器人各部分质量、关节阻尼、物体质量和摩擦等多个方面，共 95 个参数。采样后在该 episode 内保持不变，但动作时间步长和观测噪声每个时间步随机变化，模拟现实中控制器的延迟和传感器的不确定性。

**适应性策略：**

采用循环模型作为适应性策略，其中内部记忆 能总结过去状态和动作信息，帮助策略推断环境动力学。与传统方法不同，该模型无需在运行时手动识别要预测的动力学参数，可进行端到端训练，学习内部记忆的表示。

**结合 HER 和 RDPG 训练：**

借助事后经验回放（HER）处理稀疏奖励，通过重新计算目标来增加成功经验，辅助训练。采用循环确定性策略梯度（RDPG）算法进行离策略学习，训练过程中，从回放缓冲区采样数据，计算价值函数和策略梯度，更新策略和价值函数参数。

**直接提高策略质量方法：**

在论文《Zero-Shot Skill Composition and Simulation-to-Real Transfer by Learning Task Representations》中提出一种结合任务表示学习、仿真到现实训练和模型预测控制的方法，旨在高效地让机器人在无需额外训练的情况下获取执行未知任务的策略，以解决现实差距问题，实现从仿真到现实的迁移。

方法概述：先在仿真环境中利用强化学习和变分推断学习任务合适的原语技能的连续参数化（技能嵌入或潜在表示）以及基于该表示的单一策略。接着将多技能策略直接迁移到真实机器人上，在执行未知任务时，借助模型预测控制选择技能潜在向量序列驱动策略，通过仿真评估未来动作结果，从而控制真实机器人完成任务。

技能嵌入算法：预定义一组低级技能及其奖励函数，在学习联合低级技能策略的同时，学习一个嵌入函数，用潜在变量 z 对低级技能库进行参数化。通过随机嵌入函数将技能 ID 转换为潜在向量，并将其输入策略。为辅助学习，还学习一个推理函数，根据状态轨迹窗口预测输入策略的潜在向量，定义增强奖励鼓励策略为不同潜在向量生成不同轨迹，并添加策略熵奖励确保策略不收敛于单一解。

技能嵌入准则：约束嵌入分布满足高熵和可识别性两个重要属性。高熵使每个任务能诱导尽可能广泛的潜在向量分布，对应单一技能的多种变化；可识别性确保推理网络能根据轨迹窗口高置信度地预测输入策略的潜在向量，二者共同作用于策略训练。

基于模型预测控制的零样本适应：冻结在仿真中学习的多技能策略，使用 “组合器” 算法。在每个步骤，从潜在分布中采样候选潜在向量，将仿真环境状态设为真实环境观测状态，在仿真中执行冻结策略，根据新任务奖励函数计算每个候选潜在向量的总折扣奖励，选择奖励最高的潜在向量，根据该潜在向量采样策略控制真实环境，重复此过程直至任务完成。

**参考或科普性的文章：**

Sim to Real 迁移的领域随机化，详见参考2

参考：1·[机器人强化迁移学习指南：架设模拟和现实的桥梁-阿里云开发者社区](https://developer.aliyun.com/article/819764)

2·https://zhuanlan.zhihu.com/p/651042783?utm\_id=0