强化学习在复杂决策问题中取得了引人注目的成绩，但这些算法在获得合理的表现前经常依赖大量的数据。在一些现实问题中，往往无法为学习过程提供如此庞大的数据。强化学习在复杂问题中体现出对数据的贪婪，使其很难在采样困难的真实场景下获得较成功的应用。而解决这一问题的关键是，智能体可以利用过往的经验，结合新任务的少量的样本，获得可靠的决策性能，整个过程通过知识泛化以适应新任务。这种利用先验知识以适应小样本场景的学习方法，被称为小样本学习。本文在小样本场景分类和知识泛化过程分类后，调研了多种具有小样本潜力的强化学习方法。

小样本场景分类：

专门化

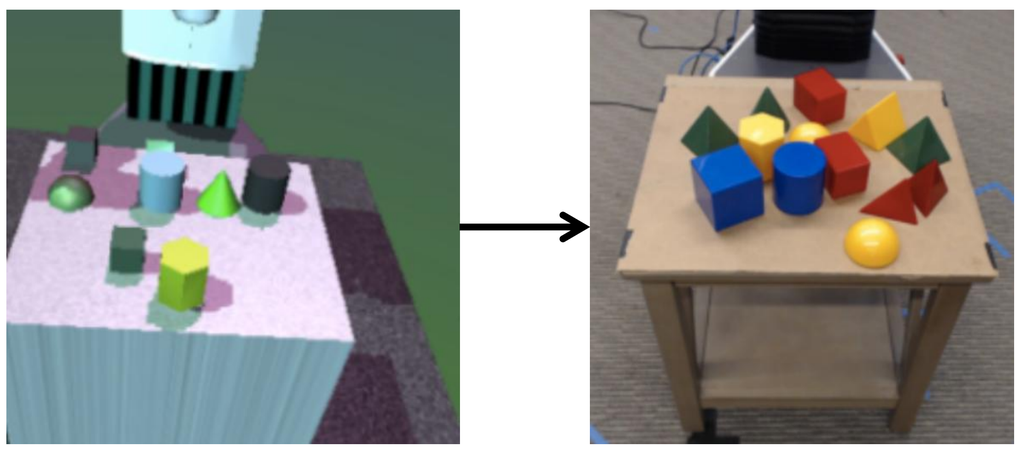


图 1. 例子：实物抓取。

将技术从模拟世界移植到真实世界（或称，*Sim2real*），可以作为一类以快速适应真实世界为目标的小样本场景，其方法为获取专门化的技术。

通用化

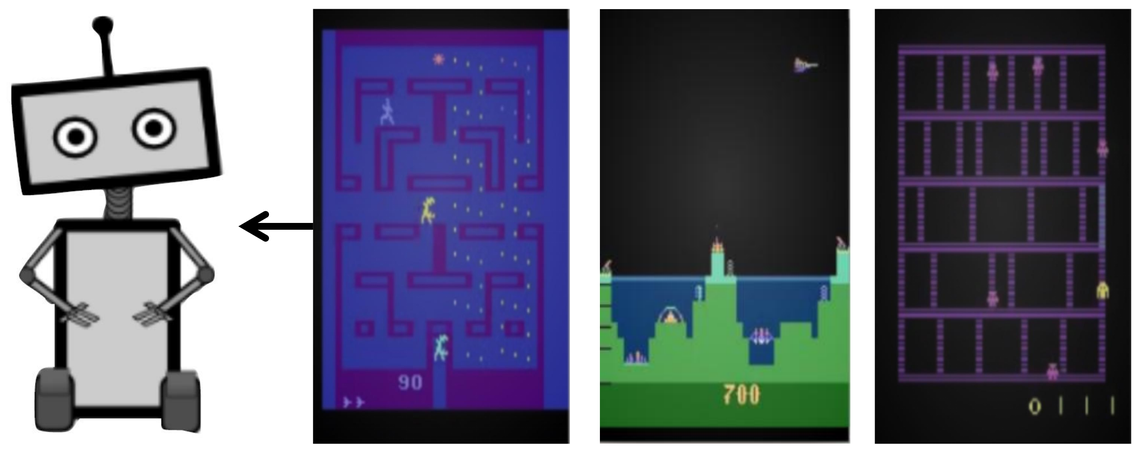
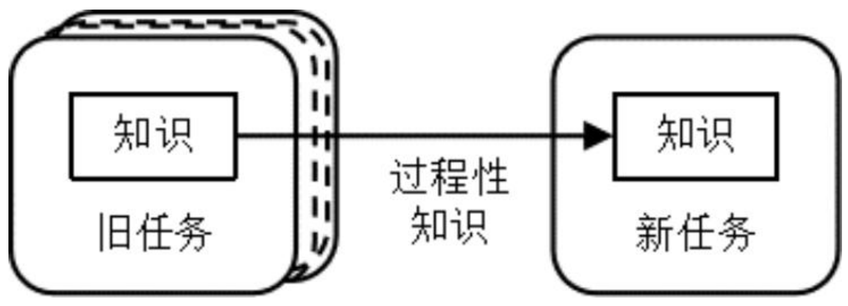


图 2. 例子：智能体玩Atari游戏。

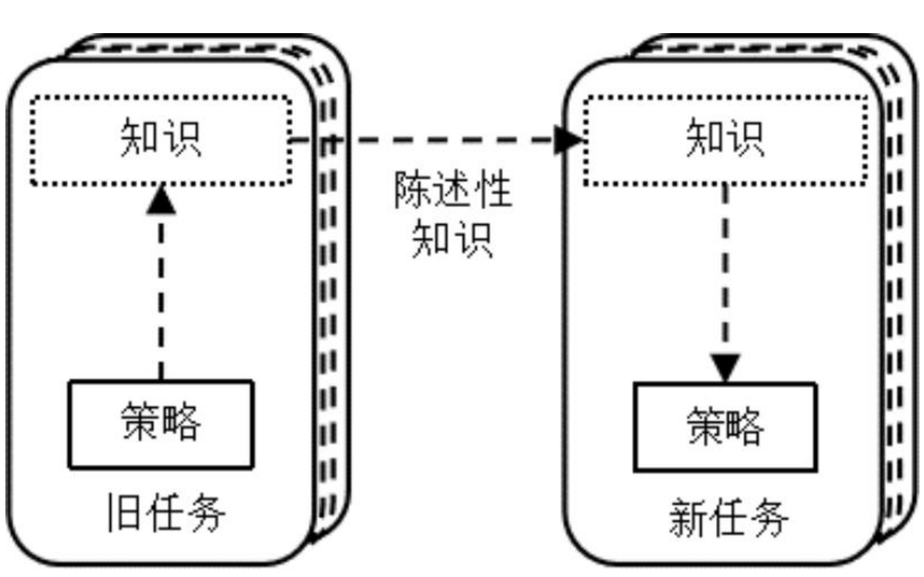
将技术在相似任务中切换，可以作为一类以快速解决同类问题为目标的小样本场景，其方法为获取通用化的技术。

知识泛化过程分类：

(a)



(b)



(c)

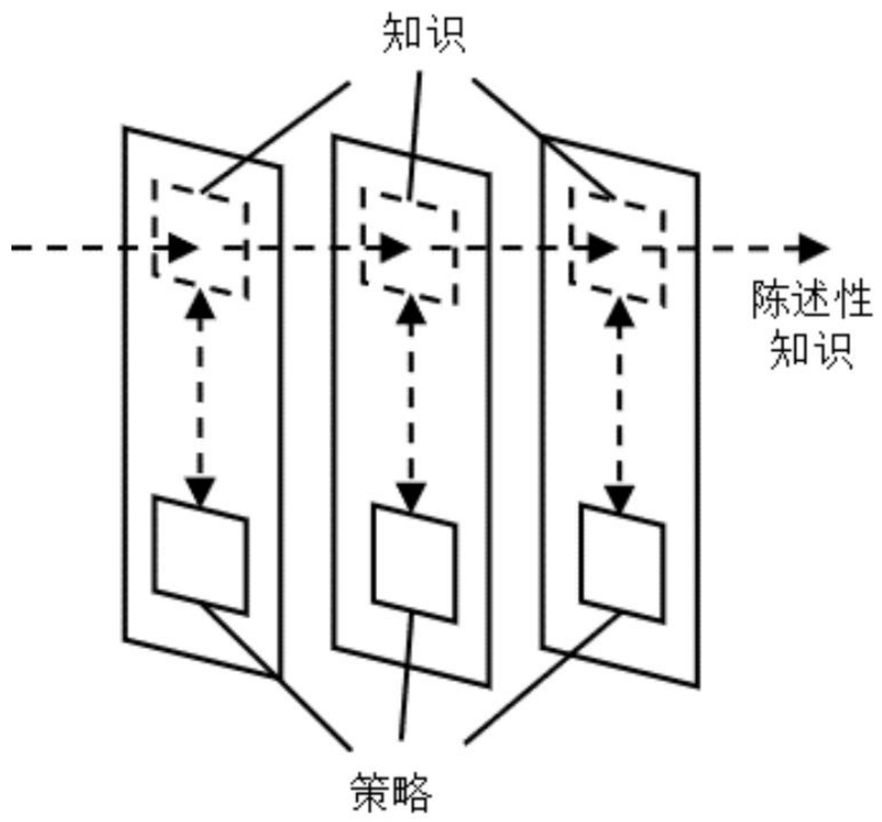


图 3. 知识泛化

如图3所示，我们将知识泛化分成两类直接的（a）与间接的（b）、（c）。“直接”和“间接”的区别是，泛化是否经过知识归纳，即是否产生了陈述性认知。对单个任务来说，智能体需要输出的是一个针对该任务的策略。若知识的主要内容是在该策略下产生的样本，那么这就是直接的。因为这些样本代表的是过程性认知。例如，样本表述了在某个状态下采取了某个动作获得多少奖赏并到达下一状态，实则反应了一个动态过程（a）。相反，若将归纳知识作为泛化内容时，即为间接的。例如，策略参数等一些不直接反应认知过程的知识（b）。这样分类的益处在于，清楚地了解了不同小样本场景下主流工作的倾向。多数工作表明直接的知识泛化更适用于在专门化的小样本场景上，而间接的知识泛化被广泛适用于通用化小样本场景。此外，就像终生学习的视角那样，间接的泛化可以被表达为如（c）所示，知识随时间在不同的任务中泛化。

我们将一些重要的算法归类如表1，在文中详细介绍各算法思想和其在小样本场景下的变现：

表 1. 本文所涉及的算法：

直接	间接
从演示中学习	元学习
样本加权	策略蒸馏
奖赏塑造	状态抽象

投稿人 姓名：王哲超 傅启明 等
单位： 苏州科技大学