摘要

在本文中，我们通过使用单个Q-table处理多个奖励来研究欺骗性强化学习的问题。强化学习的关键特征之一是与来自环境的奖励互动。这些奖励作为强化学习模型中的奖励功能存在。欺骗性强化学习可以通过向观察者展示虚假路径来将奖励功能保密。然而，在整个学习过程中仍然需要所有奖励功能信息，并且现有方法必须训练多个Q-table以实现相应的目标（Yue等，2020）。因此，我们定义了用单个Q-table处理多个奖励的问题，并设计了两个具有独特特征的模型来解决该问题。两种模型都基于强化学习和模仿（一种“掩盖事实”的欺骗性类型）。评估了两个模型并将其与其他类型的模型进行比较。结果证明，我们的模型与诚实代理相比具有欺骗性策略，与其他欺骗性强化学习相比，其时间复杂度和信息泄漏更少。

4.计算评估

在本节中，我们使用相关的路径规划框架[Masters和Sardina，2017年]测试和评估第3节中介绍的两个模型。此实验有两个目标：首先，通过比较模型来测试模型的欺骗能力 使用诚实的基准模型和先前的作品之一； 其次，将模型的路径成本与路径成本最优的诚实模型进行比较，以衡量模型的路径成本。 两项都是成功的欺骗性路径规划模型的关键指标。

4.1实验设计

在第3节中，我们介绍了两个模型：启发式模型和熵模型。 为了评估路径规划算法并可视化性能，两个模型都使用P4框架实现。 同时，将模糊模型[Yue Yang，刘正尚，Peta Masters，Tim Miller]和Astar模型作为控制组。

自变量

实验中实现的四个模型的详细信息是：

1.虚拟模型，该模型实现Astar算法以查找给定问题中的策略。 可以将此代理视为具有最低路径成本和零欺骗能力的诚实模型。

2.原始模型，它是通过使用强化学习而成功的欺骗性路径规划模型。 但是，训练了多个Q-table以处理多个奖励功能。

3.启发式模型，如第3.1节所述。

4.熵模型，如第3.2节所述。

测量

实验中引入了三种测量指标：

(1) 暴露路径相对于密度的比例。这表示：在每个密度下，已经将真实目标暴露给观察者的代理的比例。对于模型，比例由以下公式计算：

暴露路径比例=实际目标已暴露的代理数量/实验的代理总数

(2) 实际目标相对于密度的概率。这可以衡量：在每个密度下，从观察者的角度来看，真实目标成为真实目的地的可能性有多大。该度量基于幼稚的意图识别算法，通过在路径规划问题中实现成本差异的概念来进行。

（1）和（2）均用于衡量模型的欺骗能力。

(3) 成本比。在路径成本方面最优的是虚拟模型上的平均路径成本。我们希望以更低的成本获得出色的欺骗模型。

(4) 还有一个变量要提到的是“密度”。密度= x％的节点表示位置=（x％\*长度（路径））的节点。例如：对于长度为200的路径，密度为15％的代理表示它位于该路径的第30个节点。在此实验中，密度也可以视为暴露给观察者的路径的百分比。范围是0到100％，表示当密度= 100％时，代理已到达目的地。通常，模型欺骗性执行的密度越高，欺骗性就越高。

实验参数

我们在三个不同的地图上测试模型：

（1）地图没有任何障碍

（2）具有三个大障碍的地图

（3）具有一些随机但较小障碍的地图

所有三张地图的尺寸​​均相同（49 \* 49），并具有一个真实目标和多个假目标。但是，为每个地图设置了11个不同的目标位置。因此，每个模型共有33种不同的实验配置。对于度量（1）和度量（2），我们以每10％的密度记录结果。对于度量（3），仅记录平均路径成本比率。

4.2结果

图1显示了曝光路径的比例，y轴和x轴均为密度。 通常，y值越低，欺骗的能力越强。 对于此指标，很明显，原始模型表现最佳，其次是熵模型，其次是启发式模型。 从图4中可以得出三个结论：（1）尽管熵模型和启发式模型的性能不如原始模型，但它们的性能仍然比虚拟代理好得多，这证明它们具有欺骗性，但是 不同程度。 （2）对于启发式模型和熵模型，它们在路径的早期阶段都表现良好。 如图4所示，当密度<50％时，熵模型和原始模型之间的差异小于0.1。

图5表示真实目标的平均概率为y轴，而密度为x轴。 y值越小，欺骗的能力越强。 y轴= 0.5处有一个重要的检查点。由于当实际目标的概率大于0.5时，这意味着该目标大于任何其他虚假目标的概率。只要实际目标的概率大于0.5，观察者之间就不会有任何区别，观察者将始终正确地确定实际目标。通过分析y = 0.5下的结果，熵模型表现最佳，其次是原始模型和启发式模型，这与第一个度量相似。但是，在此评估指标中，启发式模型和熵模型显示出更高的竞争性能，而后一种甚至是最具欺骗性的模型。度量（1）和度量（2）的结果不一致的原因是：在每种密度下，原始模型中暴露的代理数量较少。但是，真实目标与原始模型的概率方差比熵模型之一大，这导致真实目标概率的平均值更大。

欺骗的能力不是我们问题的唯一方面，因为在假目标之间徘徊的非常偏斜的路径可能具有很大的欺骗性，但代价也很高。 因此，第三次测量揭示了每个模型的成本。 由于虚拟模型基于Astar路径规划算法，因此可实现最佳成本。 在图6中，每个模型的y轴成本比率是通过将其平均路径成本与虚拟模型的平均成本进行比较而得出的。 在我们的实验中，熵模型和原始模型共享相似的成本，而启发式模型的成本小于这两个模型。 这证明，即使启发式模型显示出较少的欺骗性，也可以在需要较低成本的情况下实施。

总体而言，四种型号的欺骗能力为（从最差到最佳）：

虚拟<启发式<熵≈原始

路径成本的等级（从最差到最好）：

熵≈原始<启发式<虚拟

5.讨论与未来工作

在本文中，我们提出了两个模型，用于隐藏来自外部观察者的强化学习中的真实奖励函数。启发式模型基于Q-learning的奖励塑造，而熵模型则通过比较每个状态的熵找到最模糊的动作。与以前的加固欺骗性路径规划方法不同，仅针对一个集成的Q-table针对多个奖励功能进行训练，这带来了更少的训练时间和更少的信息泄漏的优点。但是，熵模型的局限性在于，对假目标修剪过程实施了欧几里得距离，并且熵模型在不同地图上的性能仍然不同。因此，在将来的工作中，我们将修剪应用于价值迭代过程中，而不是探索路径。还对启发式模型进行泛化，以在不同环境中稳定运行。此外，还将在模型评估部分引入人体实验，作为对我们的计算实验的补充。