信息网络专题研究之应用层

一、文献信息

1、论文题目：Learning human behaviors from motion capture by adversarial imitation

2、作者：Josh Merel, Yuval Tassa, Dhruva TB, Sriram Srinivasan, Jay Lemmon, Ziyu Wang, Greg Wayne, Nicolas Heess

3、发表途径：ICLR 2017

4、发表时间：2017 Jul 10

二、问题意义

1、 研究背景及意义

建造可编程人形机器人的问题可以追溯到几个世纪前。从当代的角度来看，最优控制和强化学习方法使运动控制器的设计能够应对类人体的高维性，神经网络能够存储多种运动模式，这些模式可以重复使用、细化和灵活排序。

本文的目标是建立一个鲁棒（Robust 的音译，也就是健壮和强壮的意思。它也是在异常和危险情况下系统生存的能力）的程序，以构造控制器的一系列类似人类的运动，适合重用和细化时，在新的任务中使用。

然而，目前人类控制的方法不符合我们的愿望。 依赖于纯粹强化学习(RL)目标的方法往往会产生不够人性化和过于刻板的运动行为。为了从运动捕获数据中进行模仿学习，我们大量使用生成对抗性模仿学习（GAIL），这是模仿学习的一个最近的突破，以类似于生成对抗性网络的方式产生模仿策略。GAIL 的关键好处是，模仿和演示数据之间的相似性概念不必基于显式的、手工设计的度量来定义。

2、 主要研究问题

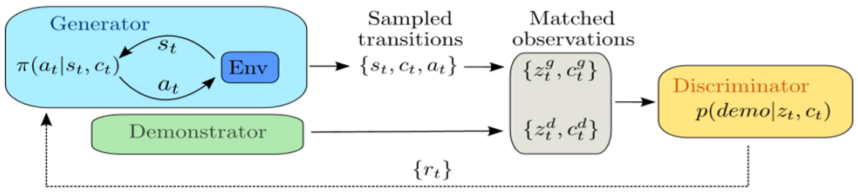
（1）对抗模仿学习（GAIL）概念及算法。

（2）基于不同场景的对抗模仿学习（GAIL）。

三、思路方法

本文先是讲解了对抗模仿学习的框架及具体的算法，为了从人体运动捕捉中完成模仿学习，我们需要证明 GAIL 即使在模仿者和演示体不同和原始控制动作未知的情况下也能工作。 我们首先使用更简单的环境来展示结果，以证明 GAIL 可以扩展到这些设置。

1、 GAIL 的框架



因为我们要学的是专家的分布，说到学分布, GAN 正好可以用在这里。Generator 就是 Actor， Demonstrator 就是专家， Discriminator 学习区分这些分布到底是专家的还是 Generator 的, Generator 要学习产生/靠近专家的分布，来骗过 Discriminator。这样 Generator

的行为就越来越像专家了. 而且因为学习的是一种分布, 就间接减小了监督学习的那种问题。

2、 GAIL 的算法

定义 Generator（Actor）产生的轨迹为$D（\tau_i）$， Demonstrator（expert）产生的轨迹为$D（\hat{\tau_i}）$。

期望找到如下式的鞍点$（\pi，D）$

$$E_{\pi_E}[\log(D(s,a))] + E_{\pi}[\log(1 - D(s,a))] - \lambda H(\pi)$$

其中π_θ是一个参数化的 policy，θ为权重。D_ω是一个参数化的鉴别器，权重为ω。

对ω使用 Adam 梯度算法，从而使上式上升。

对θ使用 TRPO 算法，从而使上式下降。

TPRO 能保证$\pi_{\theta_{i+1}}$不远离π_{θ_i}。

Generator(Actor)：产生出一个轨迹，使其与专家轨迹尽可能相近，使 Discriminator 无法区分轨迹是 Demonstrator 生成的还是 Generator 生成的。

3、 基于不同场景的对抗模仿学习

3.1 利用部分观察验证无动作的模仿

为了表明鉴别器在没有伴随动作的情况下对状态信息（包括速度）进行条件是足够的，文中验证了二维平面步行器的模仿学习，步行者的任务包括 10 个阶段，如果步行者躯干低于阈值，则提前终止。通过实验观察到比较除了由鉴别器提供的状态外的行动对模拟训练无益。在这种情况下，行动可以直接从状态的变化中推断出来。更普遍的是，加上身体和环境的形式所施加的约束，即使没有充分的驱动或决定论，也可以推断出合理的行为。

在放弃了动作之后，作者也有兴趣了解模仿在多大程度上可以跨体执行，这有时被称为“再瞄准”。作为人类观察者，我们期望不同身体的等效行为之间有一定的对应关系。通过实验发现，，模仿者可以学习使三连杆臂移动，从而使末端执行器和目标之间的向量与演示的统计量相匹配。

3.2 通过动作捕捉训练一个复杂的类人生物

初始状态分布可以手动设计成一些特定的起始姿态，具有较小的变异性。在使用 RL 训练时，我们观察到不自然的初始姿势（无论是固定的还是可变的）可以产生不同的、不自然的运动行为。 在我们期望从运动捕捉中学习的环境中，从运动捕捉数据中随机采样的姿态初始化是合理的。如果我们只是从运动捕捉姿势初始化身体，并从 RL 目标训练向前跑，我们已经可以观察到步态的自然性在视觉上的显著差异，尽管步态仍然是相当非人类的。

3.3 上下文调制和基于任务的控制

作者在一个对应于足球场一半的新环境中恢复了低级控制器，并通过键盘控制 来调节低级控制器以获得进球。 此外，作者们还考虑了是否有可能改进以前学到的技能。 他们将现有的策略和身体暴露在楼梯轨道上，并对运行策略进行微调，以通过 RL 超越楼梯，并对

前进速度进行简单的奖励。 在没有调整的情况下，运行策略很快就会落在楼梯上，但是改进后的策略能够提升和下降两个离散斜坡的楼梯，同时保持其在平坦地面上运动的能力。

最后，作者们考虑了高阶控制器对控制器的调制。高级控制器可以通过多种方法进行训练，并可以多种方式与策略交互。在这里使用一个简单的两层神经网络作为高级控制器，并训练它使用本地的、自上而下的深度摄像机向低级控制器发送上下文信号，从而实现自主导航。它的奖励对应于沿线性(3D)轨道的向前移动。

四、实验结论

GAIL 即使在模仿者和演示体不同和原始控制动作未知的情况下也能工作。

对于这些验证实验，我们首先通过 RL 训练策略，使用手工设计的奖励函数来解决简单的任务，然后根据记录的演示训练模仿策略。 这种策略有助于我们量化绩效，因为虽然我们缺乏评估 GAIL 训练的模仿策略的客观指标，但我们可以使用最初用于训练演示策略的任务目标来评估模仿策略。 然后，我们使用运动捕获构造复杂类人的策略子技能，然后在环境中的任务中探索子技能的重用。

五、启发思考

1、 GAIL 是如何结合 GAN 的思想

前面讲述 GAIL 框架时，我们直接应用了 GAN 的思想，那么 GAIL 是如何结合了 GAN 的思想的呢？在 GAN 中，我们有 Generator 和 Discriminator。其最初主要应用于图像生成，因此我们以图像生成这一应用来介绍下它的主要流程：在图像生成中，Generator 要用来学习真实图像分布从而让自身生成的图像更加真实，以骗过 Discriminator。Discriminator 则需要对接收的图片进行真假判别。在整个过程中，Generator 努力地让生成的图像更加真实，而 Discriminator 则努力地去识别出图像的真假，这个过程相当于一个二人博弈，随着时间的推移，Generator 和 Discriminator 在不断地进行对抗，最终两个网络达到了一个动态均衡：Generator 生成的图像接近于真实图像分布，而 Discriminator 识别不出真假图像，对于给定图像的预测为真的概率基本接近 0.5（相当于随机猜测类别）。

在 GAIL 中，Generator 其实就是我们的 Actor，它会根据不同的 state，采取不同的动作。而 Discriminator 将要努力区分高手的行动和 actor 的行动。对 Discriminator 来说，我们可以转化成一个简单的二分类问题，即将当前的状态和动作作为输入，得到这个动作是最优动作的概率。如果这个状态-动作对来自高手的交互样本，那么 Discriminator 希望得到的概率越接近于 1 越好，而如果这个状态-动作对来自 Generator 的交互样本，那么 Discriminator 希望得到的概率越接近于 0 越好。对 Generator 来说，我们希望自己的策略越接近于高手的策略，那么就可以使用 Discriminator 输出的概率作为奖励，来更新自身的策略，如果 Discriminator 给出的概率越高，说明我们在这一状态下采取的动作是一个较优的动作，我们就提高该动作出现的概率，反之则是一个较差的动作，降低其出现的概率。

可以看出 GAIL 的思想和 GAN 的思想如出一辙，所以 GAIL 也可以写作：GAN for Imitation Learning。