机器学习工程师纳米学位

开题报告

杜 利强 04/01 2019

项目背景

强化学习（Reinforcement Learning， RL）是与决策执行（decision making）和电机控制（motor control）有关的机器学习的子领域。它研究agent如何在复杂、不确定的环境中学习如何实现目标。强化学习让人兴奋，主要有两个原因：

- **RL非常通用，涵盖了涉及做出一系列决策的所有问题。**
  例如，控制机器人的电机，使其能够跑和跳，做出商业决策，如定价和库存管理，或玩视频游戏和棋牌游戏。RL甚至可以应用于具有顺序或结构化输出的监督学习问题。
- **RL算法已经开始在许多困难环境中取得良好效果。**
  RL有着悠久的历史，但直到深度学习的最新进展，它需要大量问题特定的工程。DeepMind的Atari结果，来自Pieter Abbeel小组的BRETT和AlphaGo都使用了深度RL算法，这些算法没有对其环境做出太多假设，因此可以应用于其他设置。

然而，RL研究也因两个因素而放缓：

- **需要更好的基准**。在监督学习中，进展由像ImageNet这样的大型标记数据集驱动。在RL中，最接近的等价物将是大量且多样化的环境集合。
  但是，RL环境的现有开源集合没有足够的多样性，并且它们通常甚至难以设置和使用。
- **出版物中使用的环境缺乏标准化。**
  问题定义中的细微差别，例如奖励函数或动作集，可以极大地改变任务的难度。这个问题使得很难重现已发表的研究并比较不同论文的结果。

四轴飞行器或**四轴飞行器无人机**在个人和专业应用领域都变得越来越热门。它易于操控，并广泛应用于各个领域，从最后一公里投递到电影摄影，从杂技表演到搜救，无所不包。

Parrot AR无人机

四轴飞行器控制器项目是强化学习唯一的毕业项目，之所以选择它，是因为它更有趣，更具挑战性，在项目中需要设计一个深度强化学习智能体，来控制几个四轴飞行器的飞行任务，包括起飞、悬停和着陆。相比而言，在其他的项目中学习器（learner）都是学会怎么做，而在RL中，智能体需要在与环境不断互动和利用的过程中选择行动以得到最大的回报。

参考：

- https://openai.com/blog/openai-gym-beta/

问题描述

这个项目旨在训练一个四轴飞行器如何飞行。具体为设计一个深度强化学习智能体，来控制几个四轴飞行器的飞行任务，包括起飞、悬停和着陆。项目的前置条件是安装ROS以及下载对应的模拟器，并且确保在模拟器运行时正确连接到ROS。项目的完成需要正确定义 **Tasks** 和 **Agents** 为每个任务。共有四个任务:

- 任务1： 起飞

  训练一个智能体成功地从地面起飞且到达某个阈值高度。
- 任务2： 悬停

  训练智能体使其起飞且悬停在设定的高度（如地上10个单位）。
- 任务3： 着陆

  训练智能体学会从地上某个位置温和地着陆。

- 任务4： 整套动作

  这个任务要求实现一个端到端的任务，即起飞，原地悬停某个时长，然后着陆。

数据集和输入

项目是一个强化学习项目，因此没有明确的数据概念，没有结果，只有目标。在这里，环境的状态和动作空间作为输入，具体为飞行器所处的位置和方向以及所受的线性推力和扭矩。

解决方案描述

如在*问题描述*中提及的，项目要求实现四个任务，需要为每个任务良好定义 **Tasks** 和 **Agents**。这里需要注意的一个问题是，动作空间是连续的，意味着只能选择适合连续状态和动作空间的算法，一种可行的方法是**深度确定性策略**，简称**DDPG**，它实际上是一种行动者-
评论者方法，但是关键在于底层的策略函数本身为确定性函数，从外部添加了一些噪点，以便采取的动作具有理想的随机性。在该项目中，将实现论文中的方法以实现飞行器的控制。

基准模型

在这里没有明确的基准模型，有的只是实现效果的对比，比如跟踪在每个阶段获取的总奖励。

评估指标

在对强化学习解决方案进行迭代时，有必要衡量智能体的效果。一个简单的跟踪指标是在每个阶段获取的总奖励。通过保存阶段统计信息到文件，以供以后分析。

项目设计

如前所述，项目本身为一个强化学习项目，并且具有连续状态和动作空间，需要选择适合的算法，在该实现中采用**DDPG**来实现智能体完成飞行器的控制任务。整个流程如下：

针对每个任务：

1. 定义任务
2. 定义智能体 （采用DDPG）
3. 运行任务

针对不同的任务优化（限制）状态和动作空间，编写阶段统计信息以跟踪智能体的效果。