齐智通讯 第 169 期 与人类合作但不需要人类数据

与人类合作但不需要人类数据

Collaborating with humans without human data (November, 24th, 2021)

齐智通讯是每周出版物，其中包含与全球人工智能对齐有关的最新内容。在此处找到所有齐智通讯资源。特别是，你可以浏览此电子表格，查看其中的所有摘要。此处的音频版本（可能尚未启用）。请注意，尽管我在 DeepMind 工作，此齐智通讯仅代表我的个人观点，而不是我雇主的观点。

强调

在没有人类数据的情况下与人类合作（DJ Strouse 等人）（由 Rohin 总结）：我们之前已经看到，如果你想在视频游戏 Overcooked 中与人类合作，它有助于针对人类模型训练深度强化学习智能体（AN #70），因此智能体“期望”与人类对战（而不是例如自己的副本，如在自我对战中）。我们可以称之为“人类感知”模型。然而，由于人类感知模型必须针对模仿人类游戏玩法的模型进行训练，因此我们需要收集人类游戏玩法数据进行训练。我们是否可以训练一个足够强大的智能体来与许多不同的智能体一起玩，包括人类作为一个特例？

本文表明，这可以通过Fictitious Co-Play (FCP) 来完成，其中我们训练我们的最终智能体对抗一群自我播放智能体及其在整个训练过程中采取的过去检查点。当在 Overcooked 中与人类合作时，此类智能体会获得显着更高的回报（相对于之前链接论文中的人类感知方法）。

在他们的消融对比中，作者发现将过去的检查点包括在你训练的人群中尤为重要。他们还测试了让 self-play 智能体具有各种架构是否有帮助，并发现它几乎没有什么区别（只要你也使用过去的检查点）。

阅读更多：相关论文：基于最大熵种群的零样本人机协调训练

> **Rohin** 的观点：你可以想象两种关于如何构建人工智能系统的不同理念——第一种选择是训练他们感兴趣的实际任务（对于 Overcooked，训练智能体与人类或人类模型对战），而第二种选择是训练一个在一些更一般的任务上更强大的智能体，希望在其中包含实际任务（本文中的方法）。除了 Overcooked，另一个例子是对一些自然语言任务的监督学习（第一哲学），与互联网 GPT 风格的预训练然后提示模型解决你感兴趣的任务（第二哲学）相比。从某种意义上说，寻求单一统一的通用人工智能系统本身就是对第二个哲学的赌注——首先你构建可以完成所有任务的通用人工智能，然后将它指向你现在想要完成的特定任务。
>
> 从历史上看，我认为人工智能主要关注第一种哲学，但近年来已经显示出第二种哲学的力量。但是，我认为问题还没有解决：第二种哲学的一个问题是，通常很难将你的系统完全“瞄准”你感兴趣的真正任务，因此它的性能不佳因为它“可能”。在 Overcooked 中，FCP 智能体不会学习可以用来提高效率的人类游戏的特定怪癖（至少在理论上，人类感知智能体可以做到）。在

自然语言中，即使你适当地提示 GPT-3，它仍然有可能最终完全漫无边际地谈论其他事情，或者忽略提及一些它“知道”但互联网上的人不会说的信息。（另见这篇文章 AN #141）

我应该指出，你还可以采用混合方法，首先使用第二种哲学训练一个大型模型，然后像第一种哲学一样根据你感兴趣的任务对其进行微调，从而获得两者的好处。

我通常对哪种方法将构建更有用的智能体感兴趣，因为这似乎与预测人工智能的未来非常相关（这反过来会影响很多事情，包括人工智能对齐计划）。

技术性人工智能对齐

学习人类意图

逆向决策建模：学习可解释的行为表示（Daniel Jarrett、Alihan Hüyük 等人）（由 Rohin 总结）：有很多工作可以从演示中学习偏好，它们在演示器身上假设的结构各不相同：例如，我们可能会认为他们是玻尔兹曼理性的（AN #12）或风险敏感的，或者我们可以尝试了解他们的偏见（**AN #59**）。本文提出了一个包含所有这些选择的框架：核心思想是将演示器建模为根据规划器选择行动；该规划器的一些参数是预先固定的，以提供对规划器结构的假设，而其他参数则是从数据中学习的。这也使他们能够将信念、决策和奖励分开，从而可以将不同的结构分别强加给每个人。

该论文提供了前向问题（如何在给定奖励的情况下计算规划器中的动作，考虑像值迭代之类的算法）和后向问题（如何计算给定演示的奖励，典型的逆强化学习设置）的数学处理. 他们在医学数据集上展示了该框架，他们在其中引入了一个具有决策灵活性、信念乐观和信念适应性参数的规划器。在这种情况下，他们指定了所需的奖赏函数，然后向后推论得出结论，就该奖赏函数而言，临床医生在诊断女性和老年患者的痴呆症时似乎明显不那么乐观。

> **Rohin** 的观点：关于这篇论文需要注意的一点是，它是一部令人难以置信的学术著作；它流畅地引用了包括人工智能安全在内的多个学科的研究，并为许多此类论文提供了有用的组织框架。如果你需要对逆强化学习进行文献综述，这篇论文是一个很好的起点。

人类的非理性：对奖励推理既有坏处也有好处（Lawrence Chan 等人）（由 Rohin 总结）：最后总结，我们看到了一个带有次优演示器的逆强化学习框架。相反，本文研究了使用次优演示器执行逆强化学习的定性效果。作者修改了贝尔曼方程的不同部分，以创建一套可能的次优演示器来研究。他们在随机 MDP 和 FrozenLake 上进行了精确推断，并在简单的自动驾驶环境上进行了近似推断，并得出结论：

1. 非理性有助于奖励推理，也就是说，如果你从非理性演示器的演示中推断出奖励（你知道非理性），你通常会比从最佳演示中推断出更多的奖励（其中你知道它们是最优的）。从概念上讲，这是因为最佳演示只告诉你最佳行为是什么，而大多数非理性也可以告诉你次优行为之间的偏好。

2. 如果你不能对非理性进行建模，你的表现可能会很差，也就是说，如果你从非理性演示器的示威中推断出奖励，但你假设演示器是玻尔兹曼理性的，那么你的表现可能会很差。

Rohin 的观点：本文与我的直觉不同的一个方式是，如果演示器实际上是系统性次优的，则假设玻尔兹曼理性的表现非常糟糕。相反，我会猜测玻尔兹曼理性会做得很好 - 不如没有错误指定的情况那么好，但只会比这更糟一点。（这就是我在论文（**AN #59**）中发现的，这对我来说很直观。）关于正在发生的事情的一些假设，主要作者同意至少是故事的一部分：

1. 在假设玻尔兹曼理性时，你推断出奖励函数的分布在激励正确行为方面“接近”正确，但分配给次优行为的奖励不同。在这种情况下，你可能会得到一个非常糟糕的日志损失（本文中使用的指标），但仍然有一个合理的策略，可以很好地获得真正的奖励（我的论文中使用的指标）。
2. 我们使用的环境可能在某些重要方面有所不同（例如，在我论文中的环境中，确定目标是最重要的，这可能比推断正确的行为或奖励要容易得多。本文使用的自动驾驶环境）。

预测

预测语言模型的进展 (Matthew Barnett)（由 Sudhanshu 总结）：这篇文章旨在预测何时可以创建“人类级别的语言模型”。为此，作者迅速涵盖了信息论和自然语言处理的基本概念，例如熵、N-gram 模型、现代语言模型和困惑。从最近最先进的模型中获得的困惑数据被收集并用于估计 - 通过线性回归 - 当我们可以预期看到未来模型得分低于某些熵水平时，接近假设的英语熵。

这些预测将跨越未来 15 年，具体取决于正在解决的数据集、方法和熵水平；有一个附有这些细节的python笔记本供好奇的读者进一步研究。作者先发制人地得出结论，“要么当前的趋势很快就会崩溃，要么人类水平的语言模型可能会在未来十年或两年内出现。”

Sudhanshu 的观点：这种快速阅读提供了基于最近结果的自然、可访问的分析，同时保持自我意识（并告知读者）潜在的改进。评论部分还包括一些有趣的辩论，例如关于 Perplexity 度量的 Goodhart 能力。

我个人觉得这些估计大致符合我自己的直觉。我什至可以说，随着文本、语音/音频、视频的改进生成能力以及多模式一致性和集成的融合，我们在大约 10 年后看到的几乎任何类型的内容都将通过算法生成，并且与人类专业人员的工作没有区别。

Rohin 的观点：我通常会采用这种方法产生的预测作为我自己的预测，也许会使其更长一些，因为我预计快速增长的计算趋势会放缓。但请注意，这是对人类语言模型的预测，而不是变革性人工智能；我预计这些会完全不同，并预测变革性的人工智能会在以后出现。

杂项(对齐)

Rohin Shah 谈 2021 年通用人工智能安全研究状况(Lucas Perry 和 Rohin Shah)(由 Rohin 总结):与往年一样(**AN #54**),在这个 FLI 播客中,我谈到了该领域的状况。相对于前几年,这个播客更具介绍性,并且更多地关注我觉得有趣的东西,而不是整个领域认为有趣的东西。

阅读更多:语音文字

推荐系统

强化学习推荐系统中的用户篡改(Charles Evans 等人)(由 Zach 总结):大规模推荐系统已经成为一种过滤大量内容以识别并向用户推荐内容的方法。然而,这些进步导致了对在应用程序中使用推荐系统的社会和伦理问题。本文重点关注使用基于强化学习的推荐系统可能带来的社会可操纵性和两极分化。特别是,他们提供的证据表明,此类推荐系统的一个重要目标是通过早期极化用户来参与用户篡改,以试图使以后的预测更容易。

为了形式化这个问题,作者引入了一个因果模型。从本质上讲,他们注意到预测用户偏好需要一个外生的、不可观察的变量来模拟点击率。然后,他们引入了工具目标的概念,该概念对基于强化学习的算法在一组潜在任务上的一般行为进行建模。作者认为,只要用户意见具有可塑性,此类算法将具有影响外生/偏好变量的工具性目标。这最终会带来偏好操纵的风险。

使用一个简单的媒体推荐问题来检验作者的假设。他们将外生变量建模为左翼、中立或右翼。用户偏好具有延展性,因为用户从对方显示的内容将极化他们的初始偏好。在实验中,作者表明,标准的 Q 学习算法将学会篡改用户偏好,从而增加左翼和右翼群体的两极分化。此外,即使智能体使用了篡改,它也无法胜过避免篡改的粗略基线策略。

> **Zach** 的观点:这篇文章很有趣,因为它形式化并通过实验证明了许多人对推荐系统的直观担忧。我还发现工具性目标的形式化具有独立的兴趣。最令人惊讶的结果是,利用篡改的智能体并不比避免篡改的策略特别有效。这表明工具性激励并没有真正指向什么是实际最优的,我发现这是一个有启发性的区别。

新闻

OpenAI 招聘软件工程师,Alignment(由 Rohin 总结):听起来确实如此:OpenAI 正在招聘一名软件工程师与 Alignment 团队合作。

BERI 招聘 ML 软件工程师 (Sawyer Bernath)（由 Rohin 总结）：BERI 正在招聘一名远程 ML 工程师，这是他们与UMass Amherst的自主学习实验室合作的一部分。目标是创建一个软件库，以便轻松部署 ALL 的 Seldonian 算法框架，以实现安全和一致的人工智能。

人工智能安全需要伟大的工程师（Andy Jones）（由 Rohin 总结）：如果前两个角色不足以说服你，这篇文章明确指出很多人工智能安全工作是优秀工程师的瓶颈，并鼓励人们申请到这样的角色。

虚拟人工智能安全营 **2022**（由 Rohin 总结）：该远程研究计划的应用程序是开放的，来自不同学科的人们聚集在一起，在已建立的人工智能研究人员的指导下研究一个开放的问题。申请截止日期为 12 月 1 日。

强化学习的政治经济学时间表（由 Rohin 总结）：NeurIPS 的PERLS 研讨会(AN #159) 的日期已定为 12 月 14 日，时间表和演讲者名单现已在网站上提供。