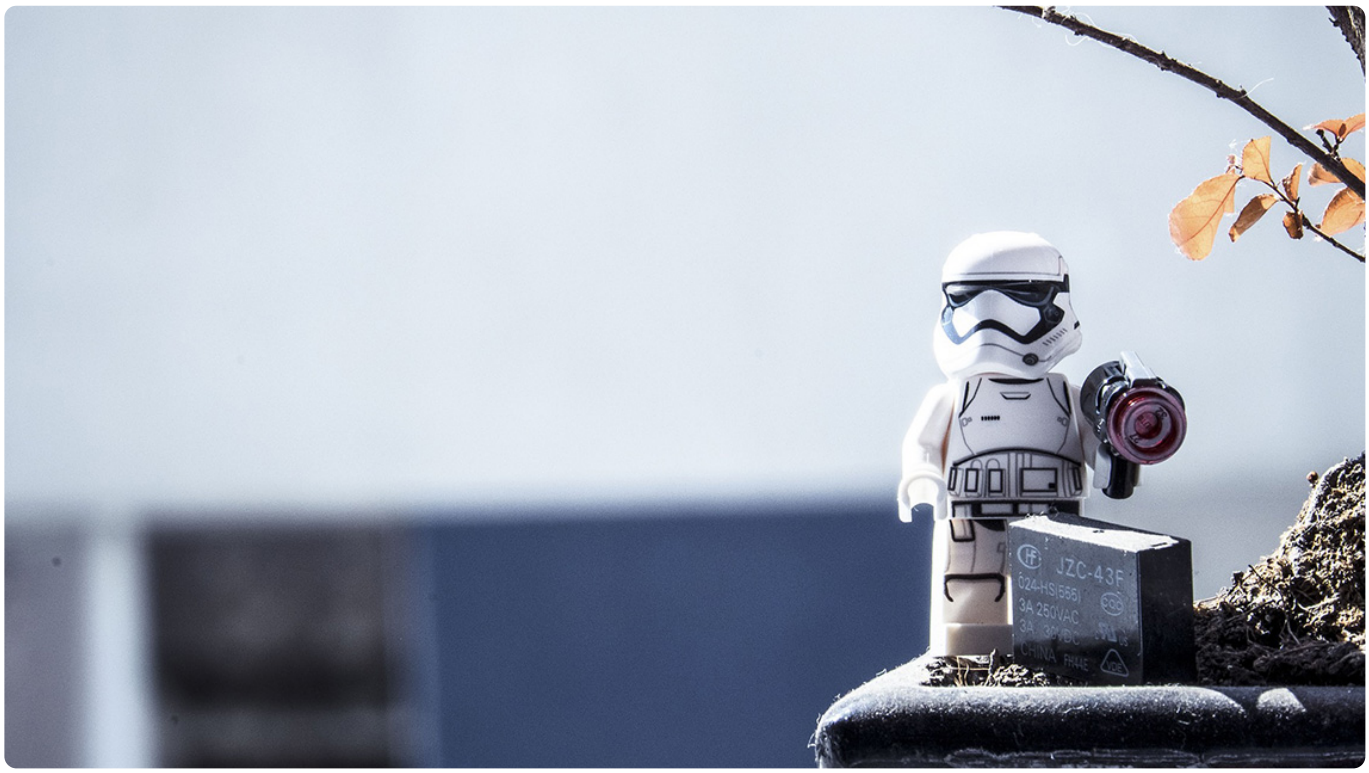
053 | 精读2017年NIPS最佳研究论文之三：如何解决非完美信息博弈问题?

2018-02-02 洪亮劼

AI技术内参 进入课程 >



讲述：初明明
时长 06:55 大小 3.17M

今天，我们来分享一下 NIPS 2017 的最后一篇最佳论文《安全和嵌套子博弈解决非完美信息博弈问题》（Safe and Nested Subgame Solving for Imperfect-Information Games）。这篇文章讲的是什么内容呢？讲的是如何解决“非完美信息的博弈”问题。

和前两篇分享的文章类似，这篇文章也是理论性很强，并不适合初学者，我们在这里仅仅对文章的主要思想进行一个高度概括。如果你对文章内容感兴趣，还是建议要阅读原文。

另外一个值得注意的现象是，即便在深度学习如日中天的今日，我们本周分享的三篇 NIPS 最佳论文均和深度学习无关。这一方面展现了深度学习并不是人工智能的全部，另一方面也让我们看到机器学习和人工智能领域的宽广。

作者群信息介绍

本文一共两位作者。

第一作者叫诺阿·布朗（Noam Brown）。布朗是卡内基梅隆大学计算机系的博士生，目前的主要研究方向是利用强化学习和博弈论的思想来解决大规模的多机器人交互的问题。这篇文章提到的“非完美信息博弈”也是这里面的一个分支问题。布朗已经在这个方向发表了多篇论文，包括三篇 AAAI 论文、两篇 NIPS 论文、一篇 ICML 论文、以及一篇 IJCAI 论文。

和本文非常相关的一个研究内容在 2017 年发表于《科学》（Science）杂志上，讲述了如何利用博弈论来解决“Heads-up 无限制扑克”（Heads-up No Limit Poker）的问题，并且在现实比赛中已经超过了人类的表现。这个工作也得到了不少媒体的报道。布朗 2017 年也在伦敦的 Google DeepMind 实习；在博士阶段之前，他曾经在金融领域工作。

第二作者是布朗的导师托马斯·桑德霍姆（Tuomas Sandholm）。桑德霍姆是卡内基梅隆大学计算机系的教授，其在“机制设计”（Mechanism Design）以及“拍卖理论”（Auction Theory）等领域有长期的研究，发表了 450 多篇学术论文，并且有超过 2 万多的引用数。除了他在学术上的造诣以外，桑德霍姆还有一些轶事，比如，他还有非常广泛的兴趣爱好，在他的主页就列举了他冲浪、喜好魔术以及对飞行的热爱。

论文的主要贡献和核心方法

我们首先来看一下这篇文章的主要贡献，弄明白这篇文章主要解决了什么场景下的问题。

对于一篇理论性很强的文章来说，我们通常需要不断地提问，这篇文章的核心主旨到底是什么，这样才能够帮助我们了解到文章的主干。

首先，文章讲的是一个“**非完美信息的博弈**”问题。这是什么意思呢？要理解“非完美信息博弈”，我们就必须要说一下“**完美信息博弈**”。

简单来说，“完美信息博弈”指的是博弈双方对目前的整个博弈状况都完全了解，对于博弈之前，以及整个博弈时候的初始状态也完全了解。在这种定义下，很多大家熟悉的游戏都是“完美信息博弈”，比如围棋、象棋等等。那么，DeepMind 开发的 AlphaGo 以及后来的 AlphaGo Zero 都是典型的针对“完美信息博弈”的人工智能算法。

“非完美信息博弈”并不是说我们不知道对方的任何信息，而只是说信息不充分。什么意思呢？比如，我们可能并不知道对手在这一轮里的动作，但我们知道对手是谁，有可能有怎样的策略或者他们的策略的收益（Payoff）等。

除了在表面定义上的区别以外，在整个问题的机构上也有不同。

“完美信息博弈”有这样的特征，那就是在某一个时刻的最优策略，往往仅需要在问题决策树当前节点的信息以及下面子树对应的所有信息，而并不需要当前节点之前的信息，以及其他的旁边节点的信息。

什么意思呢？比如我们看 AlphaGo。本质上在这样“完美信息博弈”的场景中，理论上，我们可以列出所有的棋盘和棋手博弈的可能性，然后用一个决策方案树来表达当前的决策状态。在这样的情况下，走到某一个决策状态之后，往往我们仅仅需要分析后面的状态。尽管这样的情况数目会非常巨大，但是从方法论的角度来说，并不需要引用其他的信息来做最优决策。

“非完美信息博弈”的最大特点就正好和这个相反，也就是说，每一个子问题，或者叫子博弈的最佳决策，都需要引用其他信息。而实际上，**本篇论文讲述了一个事实，那就是“非完美信息博弈”在任何一个决策点上的决策往往取决于那些根本还没有“达到”（Reach）的子博弈问题。**

在这一点上，论文其实引用了一个“掷硬币的游戏”来说明这个问题。限于篇幅，我们就不重复这个比较复杂的问题设置了，有兴趣的话可以深读论文。

但是从大体上来说，这个“掷硬币的游戏”，其核心就是想展示，两个人玩掷硬币，在回报不同，并且两个人的玩法在游戏规则上有一些关联的情况下，其中某一个玩家总可以根据情况完全改变策略，而如果后手的玩家仅仅依赖观测到先手玩家的回馈来决策，则有可能完全意识不到这种策略的改变，从而选择了并非优化的办法。这里的重点在于先后手的玩家之间因为规则的牵制，导致后手玩家无法观测到整个游戏状态，得到的信息并不能完全反应先手玩家的策略，从而引起误判。

为解决这样博弈问题，**这篇文章提出的一个核心算法就是根据当前的情况，为整个现在的情况进行一个“抽象”（Abstraction）。**这个抽象是一个小版本的博弈情况，寄希望这个抽象能够携带足够的信息。然后，我们根据这个抽象进行求解，当在求解真正的全局信息的时候，我们利用这个抽象的解来辅助我们的决策。**有时候，这个抽象又叫作“蓝**

图”（Blueprint）策略。这篇文章的核心在于如何构造这样的蓝图，以及如何利用蓝图来进行求解。

方法的实验效果

文章在“Heads-up 无限制扑克”的数据集上做了实验，并且还比较了之前在《科学》杂志上发表的叫作“利不拉图斯”（Libratus）的算法版本。人工智能算法都大幅度领先人类的玩家。

有一种算法叫“非安全子博弈算法”（Unsafe Subgame Solving），也就是说并不考虑“非完美信息的博弈”状态，把这个情况当做完美信息来做的一种算法，在很多盘游戏中均有不错的表现，但是有些时候会有非常差的结果，也就是说不能有“健壮”（Robust）的结果。这里也从实验上证明了为什么需要本文提出的一系列方法。

小结

今天我为你讲了 NIPS 2017 的第三篇最佳研究论文，文章的一个核心观点是希望能够通过构建蓝图来引导我们解决非完美信息博弈的问题，特别是在扑克上面的应用。

一起来回顾下要点：第一，我们简要介绍了这篇文章的作者群信息。第二，我们详细介绍了这篇文章要解决的问题以及贡献 。第三，我们简要地介绍了文章的实验结果 。

最后，给你留一个思考题，为什么非完美博弈的整个问题求解现在并没有依靠深度加强学习呢，大家在这个问题上有什么直观上的体会呢？

欢迎你给我留言，和我一起讨论。

上一篇　052 | 精读2017年NIPS最佳研究论文之二：KSD测试如何检验两个分布的异同?

下一篇　054 | 数据科学团队养成：电话面试指南

精选留言 (1)

写留言

林彦　2

2018-02-02

如果这个问题使用深度强化学习，感觉上这个场景是状态转移概率函数和奖赏函数都难以直接获取的免模型学习。传统的蒙特卡罗和时序差分学习都是基于采样轨迹的值来迭代更新策略。这个问题中后手能采样到的轨迹中和最优策略有可能差异会较大，这样很难生成最优策略。

...

展开