Value Iteration Networks

Aviv Tamar, Yi Wu, Garrett Thomas, Sergey Levine, and Pieter Abbeel

Dept. of Electrical Engineering and Computer Sciences, UC Berkeley

【强化学习 45】VIN

张楚珩
清华大学 交叉信息院博士在读

VIN 的全程为 value iteration network，是2016年 NIPS best paper。

原文传送门

Tamar, Aviv, et al. "Value iteration networks." Advances in Neural Information Processing Systems. 2016.

特色

本文主要设计了一种新的策略表示，这种策略表示仍然是使用神经网络，不过这里设计了新的神经网络的网络结构。这种网络结构的设计中利用了卷积神经网络（CNN）和值循环（value iteration）之间的关联，使得学习到的策略更注重长远的规划（planning），而不是只记住了状态和需要作出行动之间的单步对应关系（文章称反应式策略，reactive policy）。这是以强化学习泛化能力为目标的一个尝试。

过程

1. 动机

这里最为主要的动机是观察到值循环和卷积神经网络之间的相似之处。值循环可以写为

$$V_{n+1}(s) = \max_a Q_n(s,a) \quad \forall s, \quad \text{where} \quad Q_n(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a) V_n(s').$$

该公式可以和卷积神经网络做如下对应

- **图像的 width 和 height**：对应状态空间中不同的状态，如果是连续状态空间，应该需要做离散化；
- **输入 CNN 的 channel**：其中一个为状态价值函数 $V_n(s)$，另一些为奖励函数 $R(s,a)$，不同的离散动作对应不同的 channel，即输入 CNN 的有 $|\mathcal{A}|+1$ 个 channel；

- **CNN 的核（kernel）**：对应的是转移概率 $P(s'|s,a)$ 。通常来说，从 (s,a) 出发能够到达的状态 s' 是有限的，即只有一小部分的状态 s' 使得 $P(s'|s,a)\neq 0$ ，因此MDP 状态之间连接的局域性可以类比于 CNN 局域性；
- **MaxPooling 和 down sampling 操作**：类比于这里的 $\max$ 操作；有点不同的是公式里面写的是 $\max_a Q(s,a)$ （channel上的最大值），这里更像是 $\max_{s'\in N(s)} V(s')$ （width 和 height 上的最大值）；总之，这只是启发式地设计神经网络结构，具体里面的参数还是通过端到端（end-to-end）的算法来学习；
- **CNN 的输出**：是通过多次迭代得到的状态价值函数，用它来近似地表示最优状态价值函数 $V^*(s)$ ；
- **Parameter sharing**：值循环需要重复多次才能够收敛得到近似的最优价值函数，这里需要把设计的 CNN 网络重复多次，但是每次循环中各个参数的数值应该是一样的，这正好对应了深度学习里面的参数共享机制；
- **Attention 机制**：对应的是根据智能体所处的状态筛选出值得“注意”的价值函数值（通常是该状态及其附近的 $Q(s,a)$ ），通过选出来的这些价值函数值来最后判断并且决定智能体采取什么行动。

2. Value Iteration 模块对应新的 MDP

通过上面的类比设计出来的网络结构，虽然各个部分都代表了值循环里面的各个操作，但是其数值都是要么通过人为设计、要么通过神经网络自己学习出来的，其表示的都不是准确的值循环，因此最后得到的近似最优状态价值函数 $V^*(s)$ 也不是原问题的最优价值函数，不能直接利用这个价值函数来做出行动的选择。

假设 CNN 这一套所做的是另一个 MDP $\bar{M}=\{\bar{S},\bar{A},\bar{P},\bar{R}\}$ 上的值循环，并且转移概率函数和奖励函数都和原来 MDP 上的状态表示有函数关系 $\bar{R}=f_R(\phi(s))$ ， $\bar{P}=f_P(\phi(s))$ （这两者也通过神经网络参数来控制其函数关系），那么可以认为这个新的MDP $\bar{M}$ 和原来的MDP M 具有一定的关系，并且产生的 $\bar{V}^*(s)$ 对于原来的 MDP M 上的决策具有很大的帮助。这样我们就可以把 attention 之后选择出来的状态价值函数向量 $\psi(s)$ 作为额外的特征，传入通常的策略中进一步学习。

注意到，这整个过程都是通过可微分的网络结构来实现的，因此可以直接进行端到端的学习。

3. VIN 网络结构

通过上述分析，作者设计出了如下的网络结构。左边这个图是总体的结果，其中各个部分我们前面都已经说明了，右边是把左图中的VI Module展开绘制。

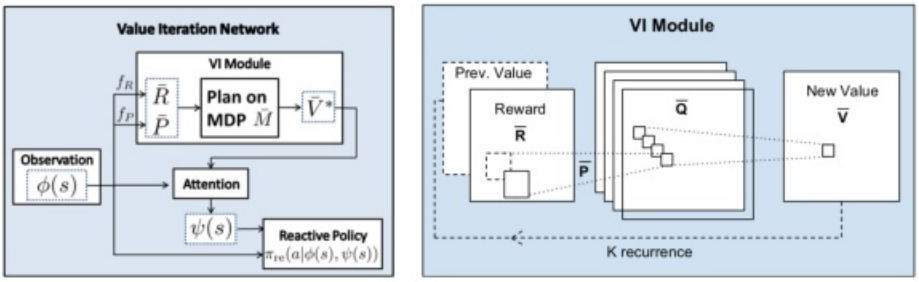

Figure 2: Planning-based NN models. Left: a general policy representation that uses value-action features from a planner to a reactive policy. Right: VI module – a CNN representation of VI algorithm.

图：VIN网络结构示意图

实验结果

文中做了四个实验：一个用于 demonstration 的 grid world 路径规划实验；一个火星表面路径规划实验；一个连续状态空间路径规划实验；一个WebNav实验。其中第二个实验没有特别有效的参照组，只与理论上界做了比较；第三个实验恰好暴露了其处理连续状态空间比较蹩脚的事实，它需要依赖连续空间的离散化；第四个实验是基于网页上的文字和查询来决定下一步进入哪个连接，目的是最后到达目标网页，这个刚好是一个建立在各个网页站点上的MDP。

我们主要来说一下第一个实验。第一个实验分为了两个部分，第一个部分的实验是 imitation learning （IL），即给出最优轨迹上的 state-aciton pair，把这个网络当做有监督学习问题来训练。通过实验结果可以看出，VIN相比于其他方法来说，单步预测的准确率差不多，但是从整条轨迹的成功率上来说更优，这说明VIN确实有一定planning的效果。

Domain	VIN			CNN			FCN		
	Prediction loss	Success rate	Traj. diff.	Pred. loss	Succ. rate	Traj. diff.	Pred. loss	Succ. rate	Traj. diff.
8×8	0.004	**99.6%**	0.001	0.02	97.9%	0.006	0.01	97.3%	0.004
16×16	0.05	**99.3%**	0.089	0.10	87.6%	0.06	0.07	88.3%	0.05
28×28	0.11	**97%**	0.086	0.13	74.2%	0.078	0.09	76.6%	0.08

Table 1: Performance on grid-world domain. Top: comparison with reactive policies. For all domain sizes, VIN networks significantly outperform standard reactive networks. Note that the performance gap increases dramatically with problem size.

第一个实验的第二部分是强化学习（RL），使用的TRPO算法来训练，这里效果更为明显，使用VIN更容易达到更高的总体成功率，并且也更容易泛化到更为困难的设定上。

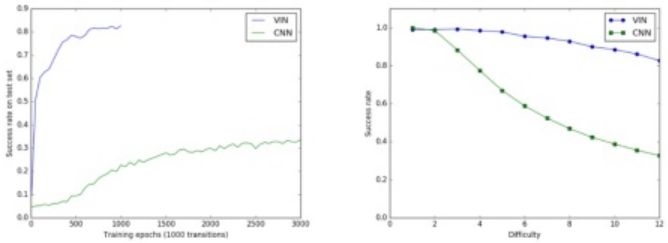

Figure 6: RL results – performance of VIN and CNN on 16×16 test maps. Left: Performance on all maps as a function of amount of training. Right: Success rate on test maps of increasing difficulty.

这里的实验结果都是在随机生成的测试集上得到的。强化学习部分使用了课程学习的技术，先学习较简单的任务，再逐步加大难度。

评述

- 本文把CNN及其相关的各种深度学习中的技术和值循环做了对应，这个创新十分有意思；
- 不过VIN的实际应用还是比较受限：
  - 首先，它要求离散的状态空间，如果状态空间是连续的就要做离散化，如果同时状态空间维度还很高，就很难再应用VIN了；（本来文章也假设了离散的行动空间，不过个人感觉连续的行动空间不会很影响VIN的效果，因为各个channel应该有一定概括能力，并且最后还会通过普通的reactive policy加工一下）
  - 其次，诚然MDP中各个状态之间是局域相连（locally connected）的，但是在大多数实际的问题中，它们之间的连接关系不可能像图像相邻像素之间的连接这样“平坦”而有规则，比如有像“虫洞”一样的连接（不恰当的比方 = =），这种情况下还能不能使用CNN来有效地做值循环就有待研究了。

PS. 之所以跑来看这篇文章是因为约了和这篇文章的作者吴翼老师（也是我院拟入职faculty）meeting，特地读一下这篇文章临时抱佛脚。

编辑于 2019-12-20

强化学习 (Reinforcement Learning)

文章被以下专栏收录

强化学习前沿
读呀读paper

进入专栏