

Relational Deep Reinforcement Learning

Vinicius Zambaldi*, David Raposo*, Adam Santoro*, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, Murray Shanahan, Victoria Langston, Razvan Pascanu, Matthew Botvinick, Oriol Vinyals, Peter Battaglia

Contact: vzambaldi@google.com, draposo@google.com, adamsantoro@google.com

DeepMind
London, United Kingdom

【强化学习算法 33】RDRL



张楚珩

清华大学 交叉信息院博士在读

19 人赞同了该文章

RDRL不是官方的名字，代表的是Relational Deep Reinforcement Learning。

原文传送门

[Zambaldi, Vinicius, et al. "Relational Deep Reinforcement Learning." arXiv preprint arXiv:1806.01830 \(2018\).](#)

特色

考虑状态空间为图像的强化学习问题（比如视频游戏），我们通常之间把图像经过CNN处理来得到低维特征。这种做法忽略了画面中不同实体相互之间的关系，因此训好这样的模型需要很多的样本，同时也只能泛化包含在训练样本分布中的情形。举例来说（不太恰当的例子），如果学会了操纵一个小兵去攻击怪物，当出现两个小兵的时候，由于智能体并不知道小兵和怪物之间的关系，因此又会把两个小兵当做另一个情形去重新学习，因此需要更多的样本；如果测试的时候出现三个小兵，智能体又会觉得这种情况没有碰到过，从而无从下手，即泛化能力不行。这篇文章就是想要教会智能体学习图像里面不同实体之间的关系，从而得到更好的sample complexity和generalization。

当然了，文章是用更为高端的语言讲的，说这是关系强化学习（relational RL），也就是要让智能体理解实体之间的关系。

文章在两类任务上面做了实验，一个是自己编的toy task叫做Box-World，另一个是SC2LE（星际2环境）里面的mini-game，都取得了不错的效果。

过程

1. 整体结构

先通过红框里面的结构把输入的一帧图像变为多个实体（entity），然后通过蓝框里面的结构学习多个实体之间的关系，最后通过绿框里面的结构把学到的东西抽象出来得到策略 π 和状态价值函数 v ；使用actor-critic方法来进行强化学习，遇到要更新策略 π 或者状态价值函数 v 的时候就反向传播通过这个结构来更新里面的参数。

注意到，如果把蓝色框换成普通的神经网络block（比如residual block）这就是一个普通的强化学

2. 如何定义图像中的实体？（红色框）

3. 如何学习实体之间的关系？（蓝色框）

$$A = \underbrace{\text{softmax} \left(\frac{QK^T}{\sqrt{d}} \right)}_{\text{attention weights}} V$$

做了上面这一大通操作，形成了一个attention block。

这样的attention block还可以叠加多层，每一层可以使用共享的参数或者不同的参数。

4. 如何形成最后的 v 和 r ? (绿色框)

attention block得到的是 $n^{n \times n \times d}$ 的矩阵，对这个矩阵做max pooling得到 n^1 的向量，然后再通过MLP即可以得到 v 和 r 了。

5. 星际环境中的特殊处理

前面讲的基本上就是Box-World环境实验中用到的结构了，对于星际环境来说，其问题结构和动作空间更复杂一些，因此需要做一些特殊处理。主要有以下几点。

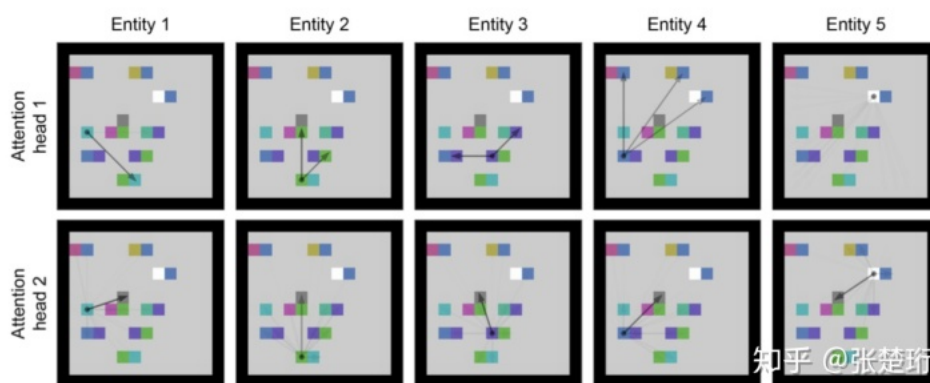
- 对于红框部分，除了使用CNN提取特征之外，还是用了LSTM让智能体能够对于最近的状态有一个认识，毕竟这个游戏从一个时间点并不能完全判断游戏的状态；
- 对于绿框部分，由于星际实验环境SC2LE的动作空间不仅包括选用一个API去调用，还需要加入一些参数，因此还需要生成这些参数。

实验

实验主要说明

- 通过在网络结构中考虑实体之间的关系，学习到的策略（渐近性能）更好；实验的对照组是把蓝框部分换成普通的residual block。
- 通过对于实体间的关系进行学习，能够有更好的泛化性能，即当画面中出现更多同样实体的时候也能够有效地处理。

这里只贴一个图，说明multi-head self-attention究竟学习到了什么。两行图分别代表两个head学到的东西，每列代表每个实体 a_i ，箭头表示该实体和其他实体之间attention的强弱，即 $a_i^T h_j$ 。可以看到，第一行里面，每个钥匙（单独的彩色块）都指向了相应的锁（两个连续块里面的右边一个，颜色和相应的钥匙相同）；第二行里面，每个钥匙都指向了小人（灰色方块）。



1. 关系 (relational) 推断什么的是很fancy的东西，我也不是很熟悉；一阶逻辑 (first-order logic) 什么的也很难理解；但是不懂这些不是很影响阅读。
2. 可能对于self-attention不够了解，可以参看下面这篇八个共同一作的文章 [Vaswani, Ashish, et al. "Attention is all you need." Advances in Neural Information Processing Systems. 2017.](#) (这篇文章非常值得一读，由于和本专栏主题不是特别相关，本专栏就不讲了)

彩蛋

网上搜到一个讲一阶逻辑的课件 (侵权，出处phil.pku.edu.cn/persona...)

4 年前为真的例子 (为了简化), 怎么让计算机理解和推理?

周迅的前男友窦鹏是窦唯的堂弟；窦唯是王菲的前老公；周迅的前男友宋宁是高原的表弟；高原是窦唯的前任老婆；周迅的前男友李亚鹏是王菲的现任老公；周迅的前男友朴树的音乐制作人是张亚东；张亚东是王菲的前老公窦唯的妹妹窦颖的前老公，也是王菲的音乐制作人；张亚东是李亚鹏前女友瞿颖的现男友。请问下列说法不正确的是：

- 王菲周迅是情敌关系；
- 瞿颖王菲是情敌关系；
- 窦颖周迅是情敌关系；
- 瞿颖周迅是情敌关系。

谓词: 前亲密关系 EX, 现亲密关系 NOW, 一种情敌关系可被定义为: $QD(x, y) := \exists z((NOW(x, z) \wedge EX(y, z)) \vee (NOW(y, z) \wedge EX(x, z)))$.

知乎 @张楚珩

4 年前的例子 (为了简化): 贵圈真乱

只能用父子关系谓词 FZ, 兄弟关系谓词 XD, 前任函数 ex, 父亲函数 father, 年纪函数 age, 以及大小谓词 <, 怎么写周迅的前男友是王菲的前老公的堂弟:

$$\exists x ((FZ(x, (ex(c_{zx}))) \wedge XD(x, father(ex(c_{wf}))) \wedge age(ex(c_{zx})) < age(ex(c_{wf})))$$

进一步可以采用更基本的性别谓词, 手足 (sibling) 谓词, 父母子女 (parent) 谓词, 年龄函数及大小谓词定义所有的关系: 如爸爸是你父母中的男性等.

知乎 @张楚珩

发布于 2018-11-29

强化学习 (Reinforcement Learning)

▲ 赞同 19 ▼

● 添加评论

🔗 分享

♥ 喜欢

★ 收藏

...

文章被以下专栏收录



强化学习前沿
读呀读paper

进入专栏