Deep Attention Recurrent Q-Network 阅读笔记

简介

15年attention机制大火，computer vision和nlp领域全都开始在模型中加入了attention机制，自然也就有人想要把attention放在强化学习中，通过注意力机制来找到agent所需要关注的图像中重点部分。

文章从DQL介绍起，然后提到了15年的模型Deep Recurrent Q-Network。DQN有一个问题在于其输入为4Frame的图像，因此其不能掌握那些需要玩家记住比4frame更远的事件的游戏。而DRQN将q-learning的最后一层改为了LSTM层并且每个step只用了最后一个frame作为dqn的输入,其好处是虽然只能看到一个frame，但是仍可以得到frame与frame之间的相关性信息。

DQN的另一个问题是其训练时间过长，一般需要12-14天的时间在GPU上训练网络。虽然有人提出了并行化的解决方案，但是并行化并不是唯一并且最有效的补救措施。

因此在上述条件下，作者提出了在DRQN中加入attention机制，命名为DARQN。想要通过这种方法来加速训练并且让DRQN可以关注到图像中的关键区域。

模型

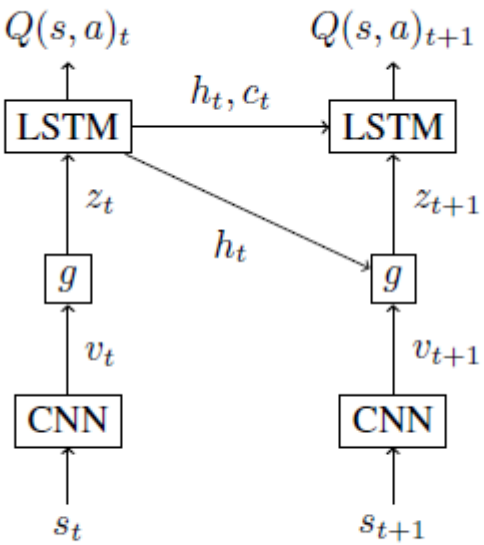

Figure 1: The Deep Attention Recurrent Q-Network

模型分为三个部分:CNN、attention、LSTM

CNN的input是84×84×1，输出是7×7的256维feature。LSTM和attention都有256的units。

attention接收到49个256维的feature之后输出其线性组合z_t,在LSTM中与h_t-1和c_t-1一起产生新的h_t。h_t用于：

1. a linear layer for evaluating Q-value of each action at that the agent can take being in state s_t
2. the attention network for generating a context vector at the next time step t+1.

接下来讲的两种attention模型就是一般attention中的两种:soft attention和hard attention在这里不做过多讲解。

attention模型的结构为两个fc layer接一个softmax layer。

$$g(v_t^i, h_{t-1}) = exp(Linear(Tanh(Linear(v_t^i) + Wh_{t-1})))/Z,$$

损失函数为：

$$J_t(\theta_t) = \mathbb{E}_{s_t, a_t \sim \rho(\cdot), r_t}[(\mathbb{E}_{s_{t+1} \sim \mathcal{E}}[Y_t \mid s_t, a_t] - Q(s_t, a_t; \theta_t))^2],$$

为了优化损失函数，作者使用了q-learning的更新规则：

$$\theta_{t+1} = \theta_t + \alpha(Y_t - Q(s_t, a_t; \theta_t))\nabla_{\theta_t} Q(s_t, a_t; \theta_t)$$

实验

作者在Breakout, Seaquest, Space Invaders, Tutankham, Gopher这几种atari游戏上进行了测试，并与DRQN、DQN进行了比较。在实验中，每100轮计算一次平均reward，一轮代表50000步。 hard和soft attention模型与DRQN一样都是按照4 unroll steps进行训练。DRQN的权重每个step都更新，而DQN和DARQN是每4step更新一次。

实验结果显示在Seaquest,Gopher中DARQN的实验结果显著好于另外两种。

	Breakout	Seaquest	S. Invaders	Tutankham	Gopher
DQN	241	1,284	916	197	1,976
DRQN	72	1,421	571	181	3,512
DARQN hard	20	3,005	558	128	2,510
DARQN soft	11	7,263	650	197	5,356

下图为实验中图片中的agent的注意点

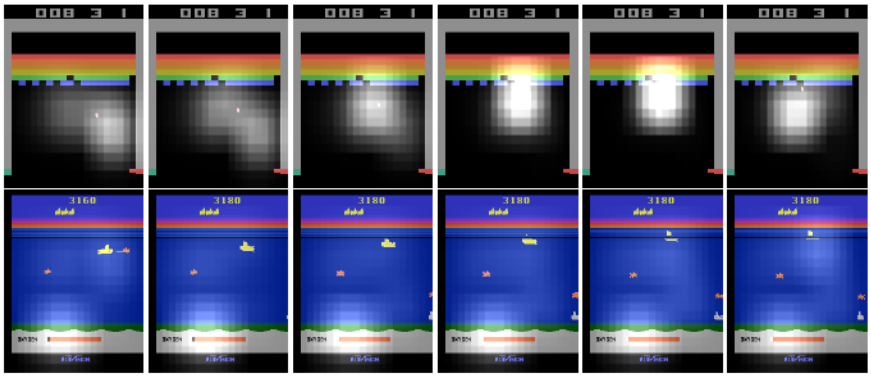

Figure 4: Visualization of attention regions for the soft DARQN model. **Top** row demonstrates the ability of the agent to focus on the ball trajectory in Breakout. **Bottom** row displays the process of submarine resurface in Seaquest. On the first screen, the agent mostly focuses on the oxygen indicator, but also notices enemies in its nearest vicinity. As the submarine rises to the surface, the attention of the agent switches to the submarine itself.

超参设置

这里值得学习的就是在进行模型的验证时，作者是用了0.05的epsilon-greedy进行25000 steps，每50000steps进行一次计算平均reward。而实验时的exploration policy使用了在1m steps中从1到0.1的线性递减epsilon的epsilon-greedy策略。

整体训练为5m steps。

目前问题和发展方向

实验中可以看出hard attention的不好大概率是策略梯度使得收敛到了局部最优的结果。

接下来的方向一方面是看能不能应用multi-scale或者glimpse的attention模型；另一方面是减少策略梯度的随机性，或者找到不同的方式训练随机注意力网络。