基于K最近邻训练的深度Q学习

2018年3月13日

1 问题描述

利用深度神经网络解决增强学习问题时，往往存在着这些问题：（1）模型需要通过激励信号学习，估计agent在一段时间内的总激励，而通常激励信号具有稀疏、有噪声和延迟的特点，即采取动作和产生激励信号之间可能间隔多步，与典型的监督学习不同；（2）深度神经网络的预设条件是样本数据相互独立，增强学习问题中的样本数据不独立，特别是相邻序列产生的状态之间高度相关；（3）增强学习中，样本数据的分布随着采取动辄的不同而发生变化，深度神经网络则假设数据样本的分布不变。

为了缓解数据样本高度相关以及分布变化带来的深度神经网络模型不稳定的情况，在模型训练过程中引入经验回放机制，将所产生的样本数据放入memory中，再随机选择一定数量的样本进行训练，能在一定程度上缓解模型的不稳定性。但是深度神经网络应用过程中存在着学习速度慢的弊病：（1）用于优化深度神经网络模型的随机梯度下降方法存在着优化速度慢的特点。随机梯度下降方法需要使用较小的学习率，如果使用较大学习率可能导致模型振荡，不收敛，而较小的学习率则导致模型收敛速度慢；（2）样本数据不平衡，低激励的样本数据数量大大超过高激励的样本数据，使得模型很难学习。为缓解上述问题，本文提出基于K最近邻训练深度Q学习方法，提高样本数据的利用效率，提高模型的训练效率。

2 算法

Q学习的最优动作-价值函数：

$$Q^*(s,a) = \mathbf{E}_{s'\sim\xi}[r + \gamma \cdot max_{a'} Q^*(s^{'}, a^{'})|s,a] \tag{1}$$

$$L_i(\theta_i) = \mathbf{E}_{s,a\sim\rho(\cdot)}[(y_i - Q(s,a;\theta_i))^2] \tag{2}$$

$$\nabla_{\theta_i} L_i(\theta_i) = \mathbf{E}_{s,a\sim\rho(\cdot);s'\sim\xi}[(y_i - Q(s,a;\theta_i))\nabla_{\theta_i} Q(s,a;\theta_i)] \tag{3}$$

基于K最近邻训练深度Q学习方法中，设置一个K最近邻字典（KND，K-nearest Neighbourhood Dict），KND支持查询和写操作，通过状态key查询到K最近邻的激励集合V_a和状态集合K_a，状态key的激励$r = \sum_i w_i v_i$，v_i是V_i的第i个最近邻对应动作a的激励，$w_i = \frac{k(h,h_i)}{\sum_j k(h,h_j)}$，$h_i$是$K_a$的第$i$个最近邻对应的状态，$k(x,y)$是向量$x$和向量$y$的核函数，如高斯核函数$k(h,h_i) = \frac{1}{||h-h_i||_2^2} + \delta$。

3 实验结果

将算法应用在Atari 2600的游戏上，结果表明相较于Prioritised Replay和Retrace(λ), 基于K最近邻训练的经验回放机制KND将相似状态的样本数据合并，得到一个更加合理的激励，基于KND的回放机制具有更高的样本利用效率，提高了模型的训练效率。