Learning Representation in Reinforcement Learning: An Information Bottleneck Approach

Chen Gong

16 December 2019

1 Conference Name

Arxiv Print，论文地址为 https://arxiv.org/abs/1911.05695。

2 Thesis statement

本文的主要目的是使用信息瓶颈框架来提高强化学习中的采样效率。首先，为什么要提高采样效率呢？因为有很多时候在强化学习中和环境的交互成本是很高的，我们希望可以让算法尽可能快的收敛，从而减少和环境的交互。也就是同样的采样量，学出来的 policy 比别的好，就是提高采样效率了。现在知道了最终的目的，那么我们为什么要使用信息瓶颈的优化来提高采样效率呢?

因为在 Atari 的游戏中，使用原始的像素作为输入，会带有很多的冗余信息来影响网络的学习。所以，我们希望通过网络的表示层来获得尽可能 “水分” 少的信息作为网络的输入，从而增加网络的分析能力。那么，我们的问题也就是如何来找一个有效的表示层。

在深度学习的实验中表明，在我们的训练过程中，神经网络首先通过使输入层和表示层变量的互信息的方式来记住输入，然后根据具体的学习任务来将把输入信息进行压缩来丢掉无用的冗余信息 (降低输入层和表示层之间的互信息)，这个过程也就是 Information E-C Process。那么，我们可以使用信息瓶颈 (IB) 的方法来加速 Information E-C Process，IB 方法可以使 RL 智能体去学习一个更有效的表示层，这个表示层会舍弃和原始输入数据不相关的数据。

3 Methods

3.1 A2C 算法基础介绍

在这里主要是分析信息瓶颈在强化学习中的运用，那么就假设大家对于强化学习有一定的基础了。在 Actor-Critic 框架中，Policy Gradient 的表达形式为:

$$\nabla_\theta \hat{J}(\theta) \approx \sum_{t=0}^{\infty} \nabla_\theta [\log \pi(a_t|X_t;\theta)(R_t - b(X_t)) + \alpha_2 H(\pi(\cdot|X_t))] = \sum_{t=0}^{\infty} \nabla_\theta \hat{J}(X_t;\theta) \tag{1}$$

其中 $R_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i}$，$H(p)$ 代表的是 p 分布的熵，$b(X_t)$ 代表的是平均函数，我们通常使用 $V^\pi(X_t)$ 来代替。实际中，完整的目标函数通常被我们表示为：

$$J(\theta) \approx \sum_{t=0}^{\infty} \log \pi(a_t|X_t;\theta)(R_t - V^\pi(X_t)) - \alpha_1 ||R_t - V^\pi(X_t))|| + \alpha_2 H(\pi(\cdot|X_t)) = \sum_{t=0}^{\infty} \hat{J}(X_t;\theta) \quad (2)$$

在表示强化学中，我们用 $J(X_t;\theta)$ 来代替 $J(Z_t;\theta)$，Z_t 是 X_t 的一个可学习的低维表示。举一个例子，给定一个表示层函数 $Z \sim P_\phi(\cdot|X)$，其中 ϕ 为参数。定义，$V^\pi(Z_t;X_t,\phi)|_{Z_t \sim P_\phi(\cdot|X_t)} = V^\pi(X_t)$。为了简化表达，我们将 $V^\pi(Z_t;X_t,\phi)|_{Z_t \sim P_\phi(\cdot|X_t)}$ 写作 $V^\pi(Z_t)$。

3.2 信息瓶颈在强化学习中的运用

3.2.1 信息瓶颈框架

我们的目的是使用信息瓶颈来使使表示层更多的丢掉冗余信息，从而加速信息提取和压缩的过程，从而使算法更快的收敛，进一步说明增加了采样效率。那么什么是信息瓶颈框架呢?

信息瓶颈是一种用来提取相关信息或者产生表示层的一种信息理论框架。那么什么是一个优秀的表示层呢？首先它可以尽可能的压缩输入变量的信息，扔掉输入信息中对于预测输出没有贡献的部分和冗余的部分。但是，也要尽可能多的可以反映输出变量之间的关系，也就是尽可能使中间变量是输出结果的充分统计量。

对于一个神经网络，我们可以写作 $X \to Y \to Z$，X 是输入变量，Z 是 X 的表示层，Y 是输出变量。而信息瓶颈就是寻找一个分布 $P(Z|X)$，使得：

$$\begin{aligned}
P^*(Z|X) &= \arg\max_{P(Z|X)} I(Y,Z) - \beta I(X,Z) \\
&= \arg\max_{P(Z|X)} \iint P(Y,Z) \log \frac{P(Y,Z)}{P(Y)P(Z)} dY dZ - I(X,Z) \\
&= \arg\max_{P(Z|X)} \iint P(Y,Z) \log \frac{P(Z)P(Y|Z)}{P(Y)P(Z)} dY dZ - I(X,Z) \\
&= \arg\max_{P(Z|X)} \iint P(Y,Z) \log \frac{P(Y|Z)}{P(Y)} dY dZ - I(X,Z) \\
&= \arg\max_{P(Z|X)} H(Y) - H(Y|Z) - \beta I(X,Z) \\
&= \arg\max_{P(Z|X)} -H(Y|Z) - \beta I(X,Z)
\end{aligned} \quad (3)$$

求解的参数与 $H(Y)$ 无关，所以被我们之间省略掉了。至于这个等式怎么来的，就需要知道互信息代表的含义，结合本小节第二段话就可以读懂了。

3.2.2 信息瓶颈结合强化学习

在强化学习中，我们怎么来使用信息瓶颈呢？根据 3.2.1 的推导，我们要得到这么一个分布使得 $-H(Y|Z) - \beta I(X,Z)$ 最大。那么，很自然，第一步就是要将 $H(Y|Z)$ 和 $\beta I(X,Z)$ 解析的表示出来。但是，$\beta I(X,Z)$ 中就包含我们已知的变量和要求的变量，所以我们可以直接进行求解，但是 $H(Y|Z)$ 中的 Y 并不是我们想要求的东西，我们需要把它解析一下。而，

$$H(Y|Z) = -\iint P(Y,Z) \log P(Y|Z) dY dZ = \mathbb{E}_{Y,Z \sim P(Y,Z)}[\log P(Y|Z)] \quad (4)$$

而在强化学习中，Y 就是累积奖励 $Y_t = R_t = \sum_{i=0}^{n} \gamma^i r_{t+i} + \gamma^{n-1} V^\pi(Z_{t+n-1})$，我们假定 $P(Y|Z)$ 服从下面分布：

$$P(Y_t|Z_t) \propto exp(-\alpha(R_t - V^\pi(Z_t))^2) \tag{5}$$

这个假设是这么来的呢？对于一个输入变量为 X_t，对应的输出变量为 Y_t，而 X_t 的表示层为 Z_t。理论是是 Y_t 和 $V^\pi(X_t)$ 直接越靠近越好，我们希望 X_t 的替代品 Z_t 自然也可以使得 Y_t 和 $V^\pi(Z_t)$ 越接近越好。所以，我们把目标分布定成 $C \cdot exp(-\alpha(R_t - V^\pi(Z_t))^2)$ 的形式。

那么我们再来理一遍逻辑，我们想使 $H(Y|Z)$ 更大，也就是使 $P(Y_t|Z_t)$ 更大，也就是使 $exp(-\alpha(R_t - V^\pi(Z_t))^2)$ 更大，也就是使得 Y_t 和 $V^\pi(Z_t)$ 越接近越好。

所以，信息瓶颈的框架可以被我们写成：

$$P^*(Z|X) = \arg\max_{P(Z|X)} \mathbb{E}_{X\sim P(X), Z\sim P(Z|X), R\sim P(R|Z)}[-\alpha(R_t - V^\pi(Z_t))^2] - \beta I(X, Z) \tag{6}$$

那么，下一步我们合并 Policy 损失 $\hat{J}(Z;\theta)$ 和 IB 损失，我们就可以得到，我们最终想要优化的损失函数为：

$$L(\theta, \phi) = \mathbb{E}_{X\sim P(X), Z\sim P_\phi(Z|X)}[\hat{J}(Z;\theta) + \mathbb{E}_{R\sim P(R|Z)}[-\alpha(R_t - V^\pi(Z_t))^2]] - \beta I(X, Z; \phi) \tag{7}$$

对比一下 (2) 中给出的 $J(Z;\theta)$ 的完整形式，我们惊奇的发现 $\hat{J}(Z;\theta) + \mathbb{E}_{R\sim P(R|Z)}[-\alpha(R_t - V^\pi(Z_t))^2] = J(Z;\theta)$，于是最终的损失函数可以被我们写为：

$$L(\theta, \phi) = \mathbb{E}_{X\sim P(X), Z\sim P_\phi(Z|X)}[J(Z;\theta)] - \beta I(X, Z; \phi) \tag{8}$$

其中，$I(X, Z; \phi)$ 表示 X 与 $Z \sim P_\phi(Z|X)$ 之间的互信息。看到这里，大家估计就会明白另一个潜在的原因，为什么结合 polic loss 和信息瓶颈了？因为他们两个的目标都是求最大值，互不冲突，可以直接加在一起。那么，使 $L(\theta, \phi)$ 最大，得到的 $P_\phi(Z|X)$ 也就是我们想要得到的那个 $P(Z|X)$，也就是：

$$P_{\phi^*}(Z|X) = \arg\max_{P_\phi(Z|X)} L(\theta, \phi) = \arg\max_{P_\phi(Z|X)} \mathbb{E}_{X\sim P(X), Z\sim P_\phi(Z|X)}[J(Z;\theta)] - \beta I(X, Z; \phi) \tag{9}$$

3.3 目标分布的导数和构造变分下界

我们的目标是：

$$\max_{p_\phi(Z|X)} \mathbb{E}_{X\sim P(X), Z\sim P_\phi(Z|X)}[J(Z;\theta)] - \beta I(X, Z; \phi) \tag{10}$$

使用 $\frac{\partial L(\theta,\phi)}{\partial P_\phi} = 0$，计算得到：

$$P_\phi(Z|X) \propto P_\phi(Z) exp\left(\frac{1}{\beta} J(Z;\theta)\right) \tag{11}$$

得到的这个之后，我们就成功的找到了极值点。但是这个极值点对应的是不是最大值呢？我们还需要进一步进行探究。

这时作者进一步证明了“表示提升理论”，这是个什么玩意呢？这个理论中证明了当目标函数为，$L(\theta, \phi) = \mathbb{E}_{X\sim P(X), Z\sim P_\phi(Z|X)}[J(Z;\theta)] - \beta I(X, Z; \phi)$ 时，如果固定住参数 θ，表示层分布 $P_\phi(Z|X)$，和状态分布 $P(X)$。如果 $P_{\hat{\phi}}(Z|X) \propto P_\phi(Z) exp(\frac{1}{\beta} J(Z;\theta))$ 时，那么就与 $L(\theta, \hat{\phi}) \geq L(\theta, \phi)$。这样就成功的得到了，公式 (11) 得到的分布就是我们想求的最优分布。

但是，问题马上又来了。尽管我们得到了最后的目标分布，但是计算 $P_\phi(Z)$ 仍然非常的复杂。为了解决这个问题，我们构建了一个变分下界分布为 $U(Z)$，我们可以使：

$$\int P_\phi(Z)\log P_\phi(Z)dZ \geq \int P_\phi(Z)\log U(Z)dZ \tag{12}$$

实际上这里我们用到了一个小小的 trick。也就是无论 $U(Z)$ 是一个什么样的分布，公式 (12) 都是必然成立的，为什么呢？因为：

$$\begin{aligned}\int P_\phi(Z)\log P_\phi(Z)dZ - \int P_\phi(Z)\log U(Z)dZ &= \int P_\phi(Z)\log\frac{P_\phi(Z)}{U(Z)}dZ \\ &= \mathbb{D}_{KL}(P_\phi(Z)||U(Z)) \geq 0\end{aligned} \tag{13}$$

那么这里的 $U(Z)$ 实际上，我们可以用任意一个分布来代替。但是，突然和你说，让你随便找反而不好找一个分布了。大多时候我们都是随意弄一个高斯分布来做，这里反而让人觉得有些奇怪。所以，我们可以得到：

$$\begin{aligned}L(\theta,\phi) =& \mathbb{E}_{X\sim P(X),Z\sim P_\phi(Z|X)}[J(Z;\theta)] - \beta I(X,Z;\phi) \\ =& \mathbb{E}_{X\sim P(X),Z\sim P_\phi(Z|X)}[J(Z;\theta) - \beta\log P_\phi(Z|X)] + \beta\int P_\phi(Z)\log P_\phi(Z)dZ \\ \geq& \mathbb{E}_{X\sim P(X),Z\sim P_\phi(Z|X)}[J(Z;\theta) - \beta\log P_\phi(Z|X)] + \beta\int P_\phi(Z)\log U(Z)dZ \\ =& \mathbb{E}_{X\sim P(X),Z\sim P_\phi(Z|X)}[J(Z;\theta) - \beta\log P_\phi(Z|X) + \beta\log U(Z)dZ] \\ =& \hat{L}(\theta,\phi)\end{aligned} \tag{14}$$

通过上述推导，我们成功的证明了 $L(\theta,\phi) \geq \hat{L}(\theta,\phi)$，这里的 $\hat{L}(\theta,\phi)$ 就是当 $P_\phi(Z) = U(Z)$ 时求得的。那么我们的目标也就变成了优化这个下界，下界提升了，自然 $L(\theta,\phi)$ 的值就变大了。所以，我们的目标分布就变成了优化最大化，

$$P_\phi(Z|X) \propto U(Z)exp\left(\frac{1}{\beta}J(Z;\theta)\right) \tag{15}$$

3.4 使用 Stein 变分梯度下降来进行优化

既然得到了我们想要目标分布,那我们想的就是如何使 $P_\phi(Z|X)$ 来靠近这个分布:$U(Z)exp\left(\frac{1}{\beta}J(Z;\theta)\right)$。我们为什么要使用 Stein 变分下降法呢？因为 Stein 变分梯度下降可以有效的解决未知归一化的情况，就如这里的公式 (14) 一样。我们这里需要进行优化的变量是 Z_i，每一步迭代的方式是：

$$Z_i \longleftarrow Z_i + \epsilon\Phi^*(Z_i) \tag{16}$$

而这里的 $\Phi^*(\cdot)$ 是粒子的分布 $P(Z)$ 和目标分布 $Q(Z)$ 之间的 KL 散度的最大下降方向 (梯度)。其中，$Q(Z) = \frac{\hat{Q}(Z)}{C}$，而 $\hat{Q}$ 是一个非归一化分布，C 是归一化参数。那么，$\Phi^*(\cdot)$ 的实际表达式为：

$$\Phi^* \longleftarrow \arg\max_{\phi\in\mathcal{H}}\left\{-\frac{d}{d\epsilon}\mathbb{D}_{KL}(P_{[\epsilon\phi]||Q}) \quad s.t. \quad ||\Phi||_{\mathcal{H}} \leq 1\right\} \tag{17}$$

而 $P_{[\epsilon\phi]}$，就是 $Z \longleftarrow Z + \epsilon\Phi^*(Z)$ 变换后的分布。那么就可以得到了梯度的方向为：

$$\Phi(Z_i) = \mathbb{E}_{Z_j\sim P}[K(Z_j,Z_i)\nabla_{\hat{Z}}\log\hat{Q}(\hat{Z})|_{\hat{Z}=Z_j} + \nabla_{\hat{Z}}K(\hat{Z},Z_i)|_{\hat{Z}=Z_j}] \tag{18}$$

其中，K 是一个核函数 (通常情况下我们使用 RBF 核函数)，我们可以注意到在 $\nabla_{\hat{Z}} \log \hat{Q}(\hat{Z})|_{\hat{Z}=Z_j}$ 式中，我们消掉了归一化因子 C，这就是 Stein 变分梯度下降法可以解决归一化不明的情况。在这个部分我们是对：

$$\mathbb{D}_{KL}(P_\phi(\cdot|X)||\frac{U(\cdot)exp(\frac{1}{\beta}J(\cdot;\theta))}{C})|_{C=\int U(Z)exp(\frac{1}{\beta}J(Z;\theta))dZ} \tag{19}$$

进行优化，实际上就是使 $\hat{L}(\theta,\phi)$ 最大化。那么我们得到的最优梯度就是：

$$\Phi(Z_i) = \mathbb{E}_{Z_j\sim P}[K(Z_j,Z_i)\nabla_{\hat{Z}}(\frac{1}{\beta}J(\hat{Z};\theta)+\log U(\hat{Z}))|_{\hat{Z}=Z_j} + \nabla_{\hat{Z}}K(\hat{Z},Z_i)|_{\hat{Z}=Z_j}] \tag{20}$$

更多有关 Stein 变分梯度下降法的理解，请读者关注 QiangLiu 的论文，Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm。而在实际的运用中，我们通常用 $\zeta \log U(\hat{Z})$ 来代替 $\log U(\hat{Z})$，来控制 $\log U(\hat{Z})$ 的大小，以防过大的情况。我们得到了梯度的方向，也就是我们已经知道了，$\frac{\partial \hat{L}(Z_i,\theta,\phi)}{\partial Z_i} \propto \Phi(Z_i)$。

但是，问题马上就来了。我们最终希望是对参数 ϕ 进行更新，那么怎么获得 ϕ 的梯度方向呢？在这里我们想了一个好办法来解决，那就是链式求导法。

$$\frac{\partial \hat{L}(Z_i,\theta,\phi)}{\partial Z_i}\frac{\partial Z_i}{\partial \phi} \propto \Phi(Z_i)\frac{\partial Z_i}{\partial \phi} \longrightarrow \frac{\partial \hat{L}(Z_i,\theta,\phi)}{\partial \phi} \propto \Phi(Z_i)\frac{\partial Z_i}{\partial \phi} \tag{21}$$

那么，我们就成功的得到了 $\hat{L}(Z_i,\theta,\phi)$ 关于 ϕ 的导数。

$$\frac{\partial \hat{L}(Z_i,\theta,\phi)}{\partial \phi} \propto \mathbb{E}_{X\sim P(X),Z_i\sim P_\phi(\cdot|X)}\left[\Phi(Z_i)\frac{\partial Z_i}{\partial \phi}\right] \tag{22}$$

而 $\phi(Z_i)$ 在我们的公式 (19) 中就已经给出了。至于关于 θ 的偏导数就非常的好求了。直接就是 policy loss：

$$\frac{\partial \hat{L}(\theta,\phi)}{\partial \theta} = \mathbb{E}_{X\sim P(X),Z_i\sim P_\phi(\cdot|X)}\left[\frac{\partial J(Z;\theta)}{\partial \theta}\right] \tag{23}$$

目标函数得到了，梯度也到了，那么我们就可以愉快的进行优化了，最后得到我们想要的带有信息瓶颈框架的强化学习算法。其他的实验细节，大家到文章的第五部分：Experiment 部分的第一段来查看吧。

4 Conclusion and Discussion

带有信息瓶颈框架的 RL 优化算法可以被我们表述为：

1. 初始化网络参数 θ,ϕ；
2. 初始化超参数 β,ζ；
3. 设置学习率：ϵ；
4. 设置从 $P_\phi(\cdot|X)$ 采样的样本数量为 M；
5. 重复下列过程：
   1) 从环境中采出一个 batch，$\{X_t,a_t,R_t,X_{t+1}\}_{t=1}^n$；
   2) 对于每个样本 $X_t\in\{X_t\}_{t=1}^n$，从 $P_\phi(\cdot|X_t)$ 中采取 M 个样本 $\{Z_i^t\}_{i=1}^M$；
   3) 获得一个 batch 的数据为 $\mathcal{D}=\{X_t,\{Z_i^t\}_{i=1}^M,a_t,R_t,X_{t+1}\}_{t=1}^n$；
   4) 计算数据 $\mathcal{D}$ 中的表示层的梯度 $\nabla_\phi L(\theta,\phi)$；

5) 计算数据 $\mathcal{D}$ 中的 RL 算法的梯度 $\nabla_\theta L(\theta, \phi)$;

6) 更新 $\phi : \phi \longleftarrow \phi + \epsilon \nabla_\phi L(\theta, \phi)$;

7) 更新 $\theta : \theta \longleftarrow \phi + \epsilon \nabla_\theta L(\theta, \phi)$;

6. 直到最后收敛为止。

这篇论文的主要贡献在于，作者观察到了 Information E-C Process 仍然存在于 RL 算法中。作者引入了信息瓶颈的框架来加速这个过程，从而加速算法的优化过程，从而提高采样的效率。信息瓶颈框架就是想得到一个表示层，尽可能的丢弃输入数据中的冗余信息，然后保留得到预测值 Y 的信息。作者通过推导，得到这个信息瓶颈框架想得到的 $P(Z|X)$ 符合一个分布，但是这个分布中的 $P_\phi(Z)$ 无法求解，于是就找了一个下界来代替它，并且对下界进行优化。作者想让这个下界去靠近一个分布，但是这个分布是一个未知归一化的过程，作者就想到了使用 Stein 变分梯度下降法来进行优化。先求得目标函数的形式，再使用 Stein 变分梯度下降法求得优化的梯度，就可以愉快的得到我们最后想要得到的模型了。除了构造下界的方法，作者还想到了可以令下界是一个均匀分布，那么在求导的时候，就可以直接省略掉了，效果也还可以。

将信息瓶颈的框架引入 RL 后，果然可以成功的加速 Information E-C Process，有效的提高采样效率。

5 My Reflections

信息瓶颈框架是个很有意思的东西，它的目标是导出一个高效的表示层。实际上看上去和 VAE 的思路有点像。但是，我们通过信息瓶颈框架得到了一个分布，这个分布是未知的，而且是不断变化的。归一化过程复杂，所以采用 Stein 变分来优化，感觉这个方法可以用的很多很多的地方，或许可以水一波。但是，还有很多数学推导的细节，值得去仔细推敲。