Contrastive Active Inference

Pietro Mazzaglia

摘要

主动推理是一个关于感知和行为的统一理论，其基础是大脑通过最小化自由能来维持世界的内部模型。从行为学的角度来看，主动推理智能体可以被看作是自我证明的生命，它的行为是为了实现其乐观的预测，即首选的结果或目标。相反，强化学习需要人类设计的奖励来完成任何期望的结果。尽管主动推理可以提供一个更自然的自我监督的控制目标，但它的适用性是有限的，因为在将该方法扩展到复杂环境方面存在缺陷。在这项工作中，我们为主动推理提出了一个对比性的目标，该目标可以极大地减少智能体学习的生成模型和规划预测模型的计算负担。我们的方法在基于图像的任务中的表现明显优于基于似然的主动推理表现得更好，同时在计算上也更便宜而且更容易训练。我们与能够获得人类设计的奖励函数的强化学习智能体进行了比较，结果表明我们的方法与它们的性能接近。最后，我们还表明对比性方法在有干扰物的情况下表现显著，而且我们的方法能够将目标推广到变化的背景环境中。

关键词：强化学习；主动推理；对比学习；自由能原理

1 引言

深度强化学习（RL）已经在多个领域取得了成功，如机器人技术。视频游戏和棋盘游戏[1-3]。从神经科学的角度来看，在基于奖励学习方面，驱动深度 RL 的奖励预测误差信号与多巴胺神经元的神经活动密切相关[4]。然而，深度 RL 中使用的奖励函数通常需要根据领域和特定任务进行人为的设计，破坏了 RL 智能体的泛化能力。此外，由于奖励函数有可能出错，致使其可能产生的意外行为，使得深度 RL 在现实世界背景下的应用存在风险。。

主动推理（AIF）最近作为一个用于统一学习感知和行动的框架出现。在 AIF 中，智能体根据一个绝对的目标进行操作：最小化他们的自由能[5]。就过去的经验而言，这鼓励更新世界的内部模型，以最大化关于感官数据的证据。关于未来的行动，推理过程变得“主动”，智能体选择满足其模型的乐观预测的行为，这些行为被表示为首选结果或目标。表示为首选结果或目标[6]。与 RL 相比，AIF 框架提供了一种更自然的编码控制目标的方式。然而，它的适用性是有限的，因为该方法在扩展到复杂环境方面时存在缺陷，而且目前的算法应用都集中在低维任务上，即侧重于低维感官输入和/或小的离散动作集的任务[7]。此外，文献中的一些实验已经将智能体的首选结果替换为奖励，从而减弱了 AIF 提供自我监督的潜力[8]。

将 AIF 扩展到具有高维度的环境，例如基于图像的环境，其中一个主要的缺陷来自于建立精确的世界模型的必要性，它试图重建感官数据中的每个细节。这种复杂性也反映在控制阶段，AIF 智能体将潜在行动的未来想象结果与他们的目标进行比较，以选择最优的行为。特别是，我们主张在图像空间中实现一个少信息量控制任务的目标。

在这项工作中，我们提出了对比主动推理（Contrastive Active Inference），这是一个 AIF 的框架，旨在通过利用对比学习，减少智能体内部模型的复杂性，并提出一个更合适的目标来实现首选结果。

我们的方法提供了一个自我监督的目标，它不断地告知智能体与其目标的距离，而不需要重建潜在行动在高维图像空间中的输出。

2 相关工作

2.1 对比学习

对比学习方法最近在无监督的学习环境中取得了重大的突破。MoCo 使用对比学习的方法，关注样本数量对学习质量影响，使用随机裁剪的样本生成方式，使得无监督学习在 ImageNet 的分类的效果超过有监督学习的性能[9]。SimCLR 通过多种数据增强的组合的方法，以及在表征和对比损失之间引入非线性变换的方式，提高了模型学习表示的质量[10]。此外，对比学习方法在用于自然语言领域[11]和无模型 RL[12]方面时也表现出了成功的效果。

2.2 基于模型的控制

基于模型的控制。由于动态生成模型的改进，使得最近基于模型的 RL 方法在控制任务和视频游戏方面上都达到了最先进的性能。Dreamer 提出利用隐状态空间进行表征学习，通过对隐状态进行想象学习外部世界的运作[13]。Kaiser 利用基于模型的强化学习算法，通过端到端的方式学习令智能体学会玩雅达利游戏[14]。另一个重要的研究方向是正确地平衡现实世界的经验与智能体内部模型产生的数据。

2.3 结果驱动控制

使用期望的结果来产生控制目标的想法在 RL 中也在进一步被探索。在[15]中，Lynch 提出了一个系统，给智能体确定一个期望的目标，智能体可以从一个潜在的空间中得出自己的行动计划，并对其进行解码，以对环境采取对应的行动。DISCERN[16]在 CNN 模型的特征空间中，使用目标和给定的观察之间的余弦相似性，使目标的相互信息最大化。Rudner[17]建立了一个新的变分推理公式，推导出一个形状良好的奖励函数，该函数直接从环境交互中学习, 从相应的变分目标中，我们还推导出一个新的概率贝尔曼备份算子，并利用它来开发一个非政策性算法来解决目标导向的任务。

2.4 主动推理

在我们的工作中，我们使用主动推理来推导行动，这只是使用 AIF 的一种可能。在其他工作中，预期自由能被动地用作效用函数，在潜在的行动序列中根据自由能选择最佳行为[18]。将神经网络的表现力与 AIF 结合起来的方法在过去几年中越来越受欢迎[19]。在[20]中，Fountas 提出了一个摊销版的蒙特卡洛树搜索，通过搜索树的方式进行行为规划，并通过学习一个习惯网络用于规划。在[21]中，AIF 被认为在小规模的任务上面，奖励最大化和探索方面的效果比传统 RL 算法表现更好。在[22]中，Millidge 提出了一个新的目标，以价值迭代的方式进行摊销规划，根据价值网络进行动作策略的学习。

3 本文方法

3.1 本文方法概述

本文采用的方法为基于对比学习的主动推理，为了减少传统主动推理对计算资源的需要以及对世界模型重建所带来的噪声影响，本文提出使用对比学习的方法改进传统自由能理论公式，基于论文给出的内容需要将主动推理的公式替换为对比学习的目标公式, 在传统主动推理算法中自由能分为外部自由能和内部自由能，而本文所采用算法则将自由能部分改写为过去对比自由能和未来对比自由能，并将模型模块分为世界模型模块以及策略价值模型模块。其中，策略模型与世界网络交互如图 1 所示：

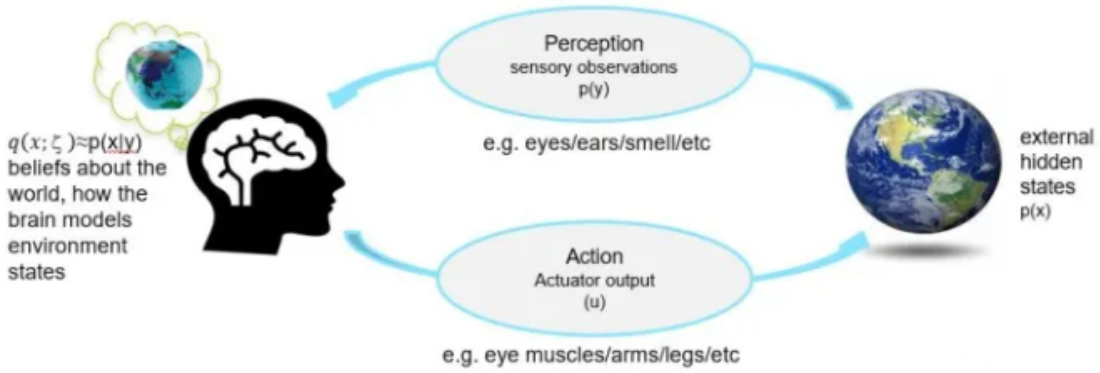

图 1: 策略模型与世界模型交互

3.2 对比主动推理损失函数定义

对比表征可通过噪声对比估计（NCE）[23]来进行学习，旨在组织数据以区分类似和不类似的数据对。根据[24]，NCE 损失可以被定义为两个变量之间互信息的下限。给定两个随机变量 X 和 Y，NCE 的下限公式可表示为

$$I(X;Y) \geqslant I_{\mathrm{NCE}}(X;Y) \triangleq \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\log\frac{e^{f(x_i,y_i)}}{\frac{1}{K}\sum_{j=1}^{K}e^{f(x_i,y_j)}}\right] \tag{1}$$

其中，期望值是来自联合分布的 K 个独立样本。$\prod_j p(xj, yj)$ 和 $f(x, y)$ 是一个函数，被称为批判函数，它用于接近于对数密度比 $log\frac{p(x|y)}{p(x)}$。最重要的是批判函数可以是无界的，来自 X 和 Y 的转换样本通过 h 和 g 函数的内积即 $f(x, y) = h(x)^T g(y)$ 会是一个很好的批判函数。

3.3 过去对比自由能损失函数定义

为了学习通过主动推理之后的环境生成模型，智能体可以通过对传统自由能公式最小化变分自由能 FAIF。但是对于高维信号，例如基于像素的图像，该模型的工作原理类似于变分自动编码器（VAE），图像在潜态 st 中编码的信息被用来通过似然模型产生高维图像的重建。

然而在像素水平上对图片进行重建有几个不足之处：(a) 它需要有大容量的模型，(b) 它的计算成本很高，(c) 有可能大部分的表示能力被浪费在与任务无关的图像的复杂细节上。

通过使用 NCE 损失可以避免预测重建后的图像。优化状态和观测之间的互信息，可以从 o_t 推断 s_t，而不必计算重建图像。为了将变分自由能损失转化为对比损失，本文将数据 $logp(o_t)$ 的常数边际对数概率加到自由能公式中，得到:

$$\begin{aligned}\mathcal{F} &\triangleq D_{\mathrm{KL}}\left[q\left(s_t \mid o_t\right) \| p\left(s_t\right)\right] - E_{q\left(s_t \mid o_t\right)}\left[\log p\left(o_t \mid s_t\right) - \log p\left(o_t\right)\right] \\ &= D_{\mathrm{KL}}\left[q\left(s_t \mid o_t\right) \| p\left(s_t\right)\right] - I\left(S_t; O_t\right)\end{aligned} \tag{2}$$

根据方程 1，我们可以引入对互信息 $I(St;Ot)$ 所适用一个下限 $I_{NCE}(St;Ot)$。对此我们可以定义过去的对比自由能为:

$$\begin{aligned}\mathcal{F}_{\text{NCE}} &= D_{\text{KL}}\left[q\left(s_t \mid o_t\right) \| p\left(s_t\right)\right] - I_{\text{NCE}}\left(S_t; O_t\right) \\ &= D_{\text{KL}}\left[q\left(s_t \mid o_t\right) \| p\left(s_t\right)\right] - \mathbb{E}_{q(s_t|o_t)p(o_t)}\left[f\left(o_t, s_t\right)\right] + \mathbb{E}_{q(s_t|o_t)p(o')}\left[\log \frac{1}{K} \sum_{j=1}^{K} e^{f(o_j, s_t)}\right]\end{aligned} \tag{3}$$

3.4 未来对比自由能损失函数定义

对行为选择进行主动推理意味着通过最小化预期自由能 G 来推断实现首选结果的行为。为了评估预期未来结果满足智能体偏好的可能性有多大，智能体使用其生成模型来预测未来的观察结果。

在对未来的图像重建预测在计算上是非常昂贵的。此外，将预测的图像与智能体在像素空间中的偏好状态相计算所包含的信息量可能很差，因为像素不应该捕获关于图像的任何语义。

当智能体学习世界的对比模型时，遵循公式 3，它可以利用拟合能力将观察结果与不重建的隐藏状态相匹配，以搜索与它的偏好相对应的状态。因此，我们将期望自由能 G 中的期望表述为首选结果，这样我们就可以添加常数边际概率 $p(o_t)$，得到:

$$\begin{aligned}\mathcal{G}\pm &= E_{\tilde{p}(o_t)q(s_t,a_t)}\left[\log q\left(s_t, a_t\right) - \log \tilde{p}\left(o_t, s_t, a_t\right) + \log \tilde{p}\left(o_t\right)\right] \\ &= D_{\text{KL}}\left[q\left(s_t\right) || p\left(s_t\right)\right] - I\left(S_t; \tilde{O}_t\right) - E_{q(s_t)}\left[\mathcal{H}\left(q\left(a_t \mid s_t\right)\right)\right]\end{aligned} \tag{4}$$

我们进一步假设 $D_{KL}[q(s_t)||p(s_t)] = 0$，它约束智能体只能修改其行动策略网络，而阻止它改变世界模型来实现其目标，这引出了未来对比自由能的以下目标函数:

$$\begin{aligned}\mathcal{G}_{\text{NCE}} &= -I_{\text{NCE}}\left(S_t; \tilde{O}_t\right) - E_{q(s_t)}\left[\mathcal{H}\left(q\left(a_t \mid s_t\right)\right)\right] \\ &= -\mathbb{E}_{q(s_t)\tilde{p}(o)}\left[f\left(\tilde{o}, s_t\right)\right] + \mathbb{E}_{q(s_t)p(o')}\left[\log \frac{1}{K} \sum_{j=1}^{K} e^{f(o_j, s_t)}\right] - E_{q(s_t)}\left[\mathcal{H}\left(q\left(a_t \mid s_t\right)\right)\right]\end{aligned} \tag{5}$$

4 复现细节

4.1 与已有开源代码对比

AIF 框架需要在统一的视图中感知和行动。在实现中，这被称为学习一个世界模型，以捕捉环境的潜在状态，最小化过去的自由能，以及学习一个行为模型，它提出行动来完成代理的偏好，最小化未来的自由能。在这项工作中，本文利用深度神经网络的高度表达能力来学习世界和行为模型。其中我根据本文所给出的网络结构使用了以下的神经网络进行学习，在代码中我增加了 VAE 编码方式作为先验训练世界模型，世界模型由以下网络组成:

Prior network(MLP,VAE): $p_\phi\left(s_t \mid s_{t-1}, a_{t-1}\right)$

Posterior network(MLP,CNN): $q_\phi\left(s_t \mid s_{t-1}, a_{t-1}, o_t\right)$

Representation model(GRU): $f_\phi(o, s)$

VAE model(MLP): $f_\phi(o|s)$ $f_\phi(s|o)$

在动作模型中我相比较于原论文采用了 VAE 作为动作模型的 Loss 项, 动作行为模型如下:

Action network(MLP,VAE): $q_\theta\left(a_t \mid s_t\right)$

Expected utility network(MLP): $g_\psi(s_t)$

其中与源代码对比增加了 VAE 网络，网络结构如下：

```python
class VAE(nn.Module):
    def __init__(self,obs_dim,z_dim):
        super(VAE, self).__init__()

        self.fc1 = nn.Linear(obs_dim, 512)
        self.fc2_mu = nn.Linear(512, z_dim)
        self.fc2_log_std = nn.Linear(512, z_dim)
        self.fc3 = nn.Linear(z_dim, 512)
        self.fc4 = nn.Linear(512, obs_dim)

    def encode(self, x):
        h1 = F.relu(self.fc1(x))
        mu = self.fc2_mu(h1)
        log_std = self.fc2_log_std(h1)
        return mu, log_std

    def decode(self, z):
        h3 = F.relu(self.fc3(z))
        recon = torch.sigmoid(self.fc4(h3))
        return recon
```

图 2: VAE 模型

4.2 实验环境搭建

本实验基于 linux 系统环境进行实验，实验环境搭建步骤如下：

①linux 中安装 Anaconda 进行环境配置，创建 python 虚拟环境

②配置算法所需依赖包，根据显卡 cuda 版本，安装相应 Pytorch 版本

③配置仿真环境，安装 Open-AI 仿真环境 gym，并安装特定实验环境 dm-controller

④配置 tensorboard 环境，记录实验数据，安装数据可视化环境

实验硬件、软件环境如下：

表 1: 硬件环境

类别	型号
CPU	Intel(R) Xeon(R) Gold 5320 CPU @ 2.20GHz
内存	金士顿 3200MHz 16GB
硬盘	三星 870 EVO 1TB
显卡	GeForce RTX 3090

表 2: 软件环境

名称	版本
Linux	Ubuntu 20.04.5 LTS
pytorch	1.7.1-GPU
cuda	11.0
gym	0.18.0
dm-controller	1.0.8

4.3 创新点

针对实验复现结果，发现实验存在问题，文章源码在运行到后期时，模型会发生过拟合情况导致世界模型以及策略模型在进行策略选取时导致结果发生剧烈震荡，致使模型的策略选择出现错误，模型所控制的机械臂不能到达预期目标。

针对源码所产生的问题进行分析，发现在源码存在以下方面的不足：

①l 代码编码方式，原文采用硬编码方式对隐状态空间进行编码，即初始状态编码为全零矩阵，通过 GRU 网络的学习能力对后续状态进行编码，但这会使得连续观察之间的编码不存在向量空间关系，从而使得下游网络出现过拟合状况。

②策略网络中的对比目标 loss 存在缺陷，由于各隐状态之间由 GRU 网络学出，隐状态之间可能会十分相似，并且 loss 中完全根据对比目标进行学习，缺少正则项约束，导致策略网络在隐状态相似情况不能很好收敛或出现过拟合状况。

针对上述问题，我提出了使用 VAE 编码器进行软编码，通过软编码将数据送入世界模型中，使得世界模型无需再学习隐状态之间的空间表示，而只关注于每个网络自己的任务，增强网络的学习能力，同时将学习出来的 VAE 模型 loss 作为策略网络的约束项，防止动作策略模型仅依靠对比表征来进行学习，CNN 对于类似图像的特征提取会导致局部相似，导致不同的状态下所取得的得分会相似，使用 VAE 隐状态编码对其进行区分，并对策略选取进行约束。

其中 VAE 网络结构如图所示：

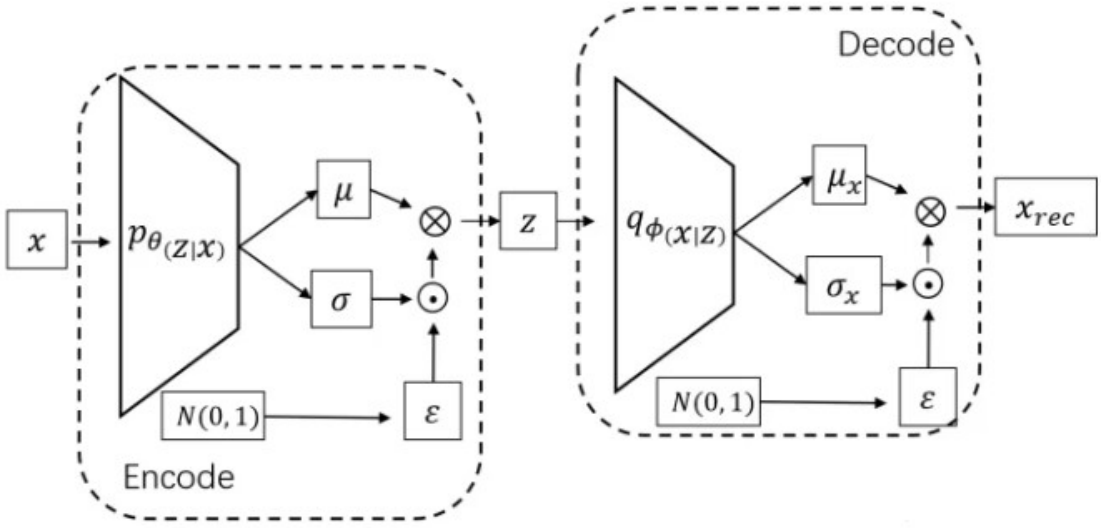

图 3: VAE 模型

5 实验结果分析

5.1 环境仿真

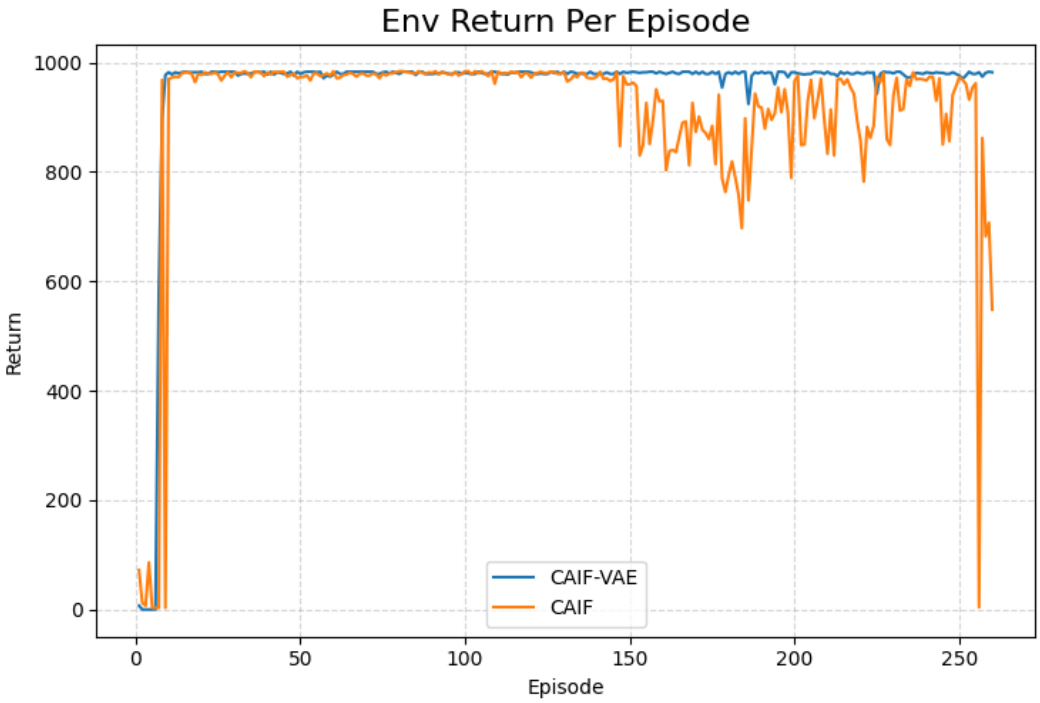

图 4: Episode Return

对整体实验结果进行分析如图 4 所示，在添加了 VAE 模型进行编码以及增加了策略网络的约束下，CAIF-VAE 在实验仿真中学习的速度性能与 CAIF 相当，在模型探索阶段，CAIF-VAF 在每个 Episode 中获取的 Return 较 CAIF 平稳，并在得分上超过 CAIF。

5.2 策略模型

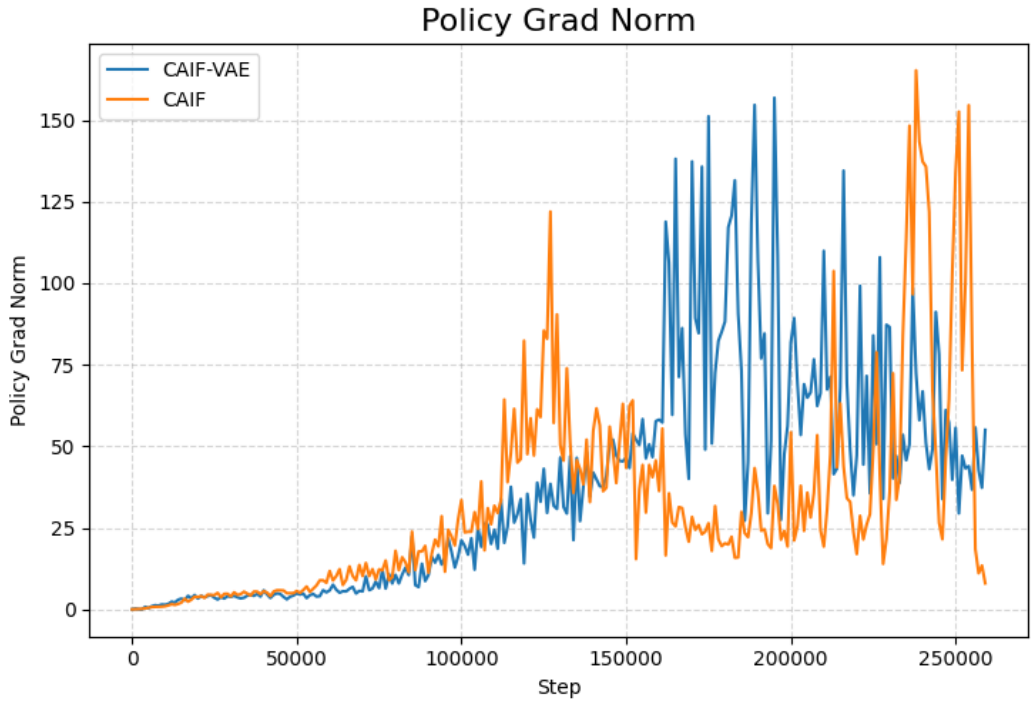

图 5: Policy Grad Norm

策略网络梯度如图 5 所示，CAIF-VAE 的策略网络由 VAE 作为正则项进行训练，从图中可以看 VAE 在网络中起到了约束作用，CAIF-VAE 随着训练轮次的增加，梯度逐步平稳下降，而 CAIF 在训练轮次较后的情况下梯度会产生较大偏差，这是由于策略网络在面对缺乏编码空间关系下的隐状态时的泛化性较低。

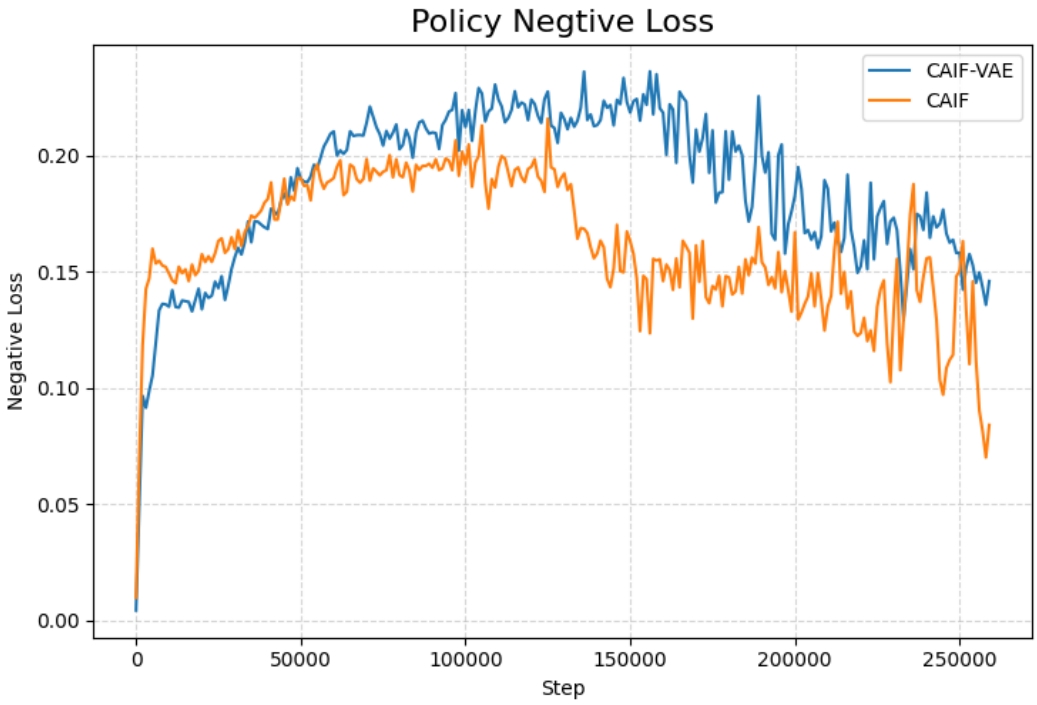

图 6: Policy Negtive Loss

策略网络针对预期状态下的 Loss，即当前状态对未来的预测与目标状态之间的互信息得分如图 6 所示，CAIF-VAE 算法相比 CAIF 在保持互信息得分的情况下较为稳定，同时使其保持在较高的水平，CAIF 在随着轮次增加的情况下模型会趋向不稳定，预测出的互信息有较大波动且相对较低。

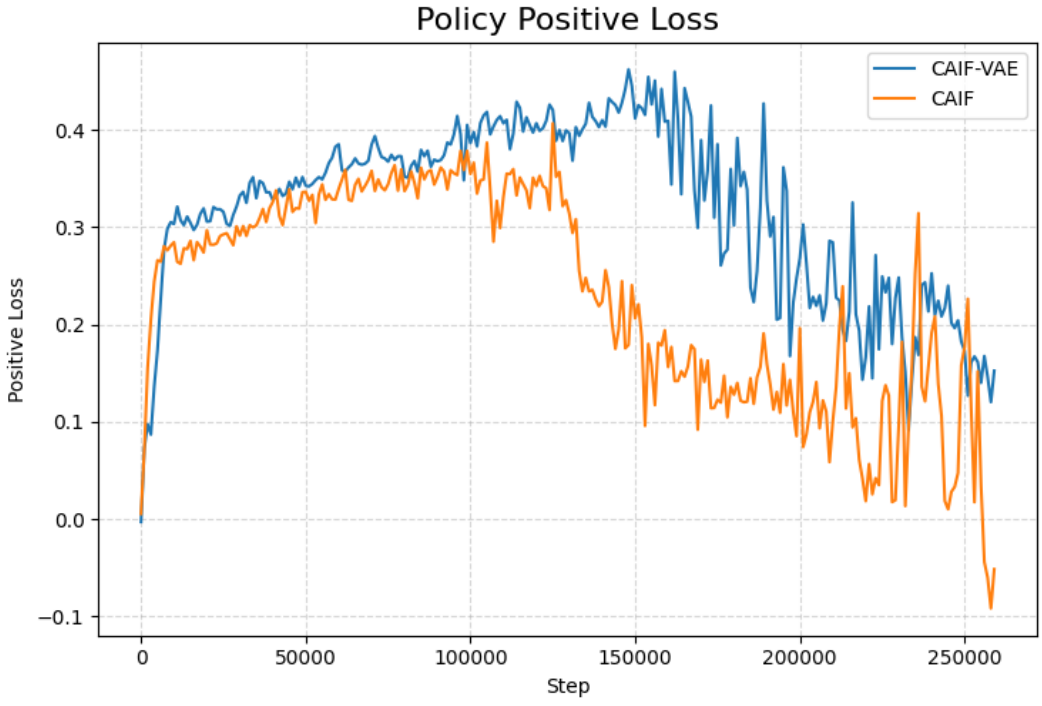

图 7: Policy Positive Loss

策略网络针对隐状态置信度 Loss，即当前状态与当前观测互信息 Loss 如图 7 所示，CAIF-VAE 算法相比 CAIF 在下降趋势方面较为稳定，但是在下降速度方面较为缓慢，这是由于增加了 VAE 作为约束项和编码针对隐状态的学习更平稳。

5.3 世界模型

图 8: World Model Image Loss

世界网络 GRU Loss 如图 8 所示，GRU 网络主要用于预测未来隐状态，CAIF-VAE 算法相比 CAIF 性能相差不多，但由于在 CAIF-VAE 中 GRU 需要学习 VAE 中的空间关系可能导致的 Loss 较高。

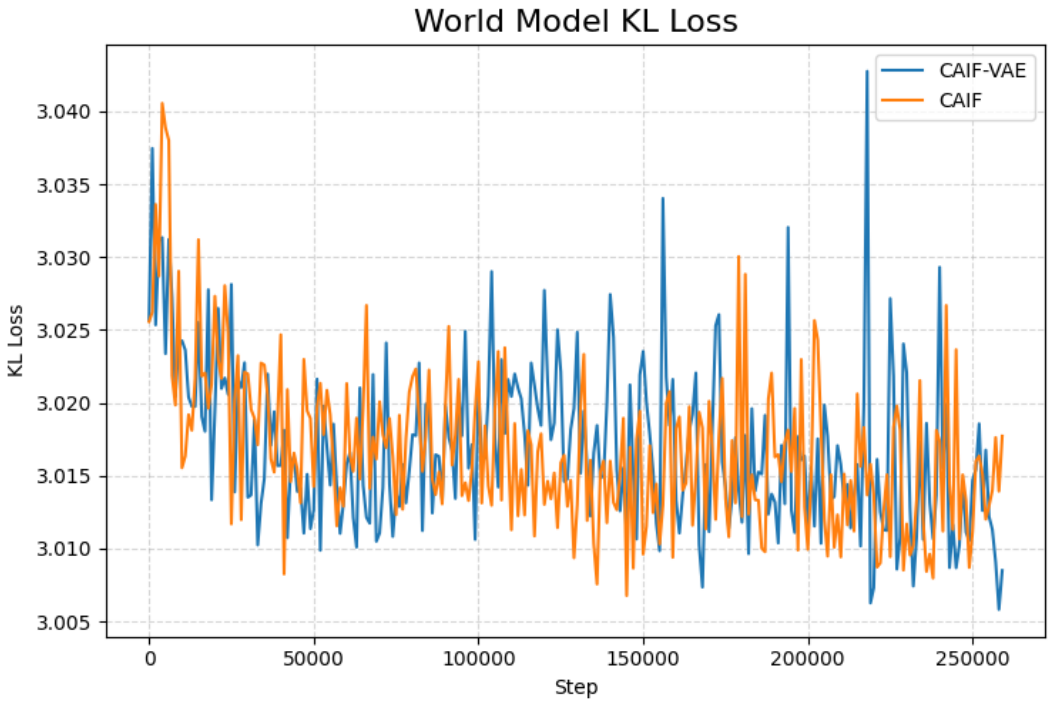

图 9: World Model KL Loss

世界网络探索约束项 KL Loss 如图 8 所示，该 Loss 主要进行约束同时使模型增加探索能力，CAIF-

VAE 算法相比 CAIF 的 Loss 波动较大，但总体趋势仍在下降，原因是由于 VAE 编码学习成本较大，导致模型波动较大。

6 总结与展望

总的来说，作者所提出的算法是目前主动推理领域较为创新的方法，应用在强化学习领域也取得了不错的效果，本文所提出的算法追求轻量化，训练速度、参数计算量都远比传统主动推理算法优秀，而我的改进增加 VAE 编码器，防止了策略模型过拟合的状况，并对隐状态进行编码提供向量空间上的连续关系，取得了不错的效果，但是由于增加了网络模型即增加了参数量，并在过程中多次使用 VAE 进行编码解码，导致在模型训练速度和参数轻量化方面远不如原论文，后续工作希望能够使用更优秀方法兼顾模型的性能以及效率。

参考文献

[1] SCHRITTWIESER J, ANTONOGLOU I, HUBERT T, et al. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model[J]., 2019.

[2] OpenAI, AKKAYA I, ANDRYCHOWICZ M, et al. Solving Rubik's Cube with a Robot Hand[Z]. 2019.

[3] BADIA A P, PIOT B, KAPTUROWSKI S, et al. Agent57: Outperforming the Atari Human Benchmark [Z]. 2020.

[4] BOGACZ R. Dopamine role in learning and action inference[J]., 2019.

[5] FRISTON K, FITZGERALD T, RIGOLI F, et al. Active inference and learning[J]. Neuroscience & Biobehavioral Reviews, 2016, 68: 862-879.

[6] FRISTON K J, DAUNIZEAU J, KILNER J, et al. Action and behavior: a free-energy formulation[J]. Biological Cybernetics, 2010, 102(3): p.227-260.

[7] LDCA B, TP B, NS B, et al. Active inference on discrete state-spaces: A synthesis[Z].

[8] FOUNTAS Z, SAJID N, MEDIANO P, et al. Deep active inference agents using Monte-Carlo methods [Z]. 2020.

[9] HE K, FAN H, WU Y, et al. Momentum Contrast for Unsupervised Visual Representation Learning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020.

[10] CHEN T, KORNBLITH S, NOROUZI M, et al. A Simple Framework for Contrastive Learning of Visual Representations[J]., 2020.

[11] OORD A, LI Y, VINYALS O. Representation Learning with Contrastive Predictive Coding[J]., 2018.

[12] SRINIVAS A, LASKIN M, ABBEEL P. CURL: Contrastive Unsupervised Representations for Reinforcement Learning[J]., 2020.

[13] HAFNER D, LILLICRAP T, BA J, et al. Dream to Control: Learning Behaviors by Latent Imagination [C]//International Conference on Learning Representations. 2020.

[14] KAISER L, BABAEIZADEH M, MILOS P, et al. Model-Based Reinforcement Learning for Atari[J]., 2019.

[15] LYNCH C, KHANSARI M, XIAO T, et al. Learning Latent Plans from Play[J]., 2019.

[16] WARDE-FARLEY D, TOM V, KULKARNI T, et al. Unsupervised Control Through Non-Parametric Discriminative Rewards[Z]. 2018.

[17] RUDNER T, PONG V H, MCALLISTER R, et al. Outcome-Driven Reinforcement Learning via Variational Inference[Z]. 2021.

[18] FRISTON K, COSTA L D, HAFNER D, et al. Sophisticated Inference[J]. Neural Computation, 2021, 33(3): 713-763.

[19] CATAL O, VERBELEN T, NAUTA J, et al. Learning Perception and Planning With Deep Active Inference[J]. IEEE, 2020.

[20] FOUNTAS Z, SAJID N, MEDIANO P, et al. Deep active inference agents using Monte-Carlo methods [Z]. 2020.

[21] TSCHANTZ A, MILLIDGE B, SETH A K, et al. Reinforcement Learning through Active Inference[J]., 2020.

[22] MILLIDGE B. Deep Active Inference as Variational Policy Gradients[Z]. 2019.

[23] GUTMANN M, HYVÄRINEN A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models[J]. Journal of Machine Learning Research, 2010, 9: 297-304.

[24] POOLE B, OZAIR S, OORD A, et al. On Variational Bounds of Mutual Information[J]., 2019.