Use tricks of GAN in Actor-Critic methods
在强化学习的AC框架中使用来自对抗网络的技巧

First Author
Institution1
author1@i1.org

Second Author
Institution2
author2@i2.org

Abstract

1. Writeup

简单思路：

把对抗网络视为强化学习AC框架的一种特殊情况。然后将对抗网络中的一些结构用于改进AC框架，这些结构有：谱归一化（WGAN）、Two time scale update rule、判别器的缓存结构。另外，今年来在深度学习领域得到广泛应用的网络结构也被尝试在AC中使用，如：BatchNorm、Dropout、DenseNet、SE-Net（Squeeze-Excitation Network-s）。然后在OpenAI gym 的二维环境Box2D 以及三维环境MuJoCo 中的机器人动作控制任务中进行测试与评估，比较训练速度与训练效果。

完整思路：

对抗网络与强化学习的AC框架都是使用了深度神经网络的多层优化问题（multilevel optimization with deep networks.）。如图，通过分析对抗网络与强化学习的 AC框架的数据流结构，我们可以发现它们的结构非常相似，并且对抗网络可以被视为强化学习AC框架的一种特殊情况。

因此，猜想：在对抗网络被使用的一些训练技巧和模型结构可以适当地通用，解决强化学习AC框架中存在的问题，反之亦可。

于是我们将对抗网络与AC框架出现过的问题与对应的解决方案进行整理。得到：

模型震荡 Model Oscillations 在对抗网络的训练后期，有时候可以观察到生成器生成的图片质量出现周期性的变化。当判别器的损失会突然增大，接着生成数据的质量明显下降。随着训练的进行，判别器的损失会慢慢变小，生成数据的质量也会慢慢恢复。而后判别器的损失又突然地增大，呈现周期性的规律。这可能是由于判别器在训练的时候无法为生成器提供适合的梯度，导致生成器往错误的方向优化，因而生成数据的质量下降。

- **双时间梯度更新规则 TTUR Two-time-scale Update Rule** 每更新一次生成器，它会更新多次判别器，并且生成器与判别器使用不同的学习率进行训练。这个解决方案认为以上问题出现的原因是：判别器是因为训练不足，无法为生成器提供了正确的梯度，或者是生成器的学习步长超过了判别器的信任域。
- **判别器缓存 Buffer of Discriminator** 将生成器的输出更新到缓存中，在训练判别器的时候，取出这些历史历史数据对判别器进行的额外训练。这个方案认为：判别器缺乏对历史生成图片的训练，导致其泛化能力下降，无法为生成器提供正确的梯度。
- **参数历史平均 Historical Averaging** 不更新目标网络的参数，而是将目前的网络参数与历史参数加权平均后得到目标网络参数。与AC框架使用的软更新（soft update）完全一致。

模式崩塌 Mode Collapse 对抗网络最后的生成结果只有少数几个模式，即便对生成器的输入进行调整，也无法得到多样性的生成结果。

- **Wasserstein GAN** 通过引入 Wasserstein Distance 使得判别器满足限制导数小于K的L连续（K-Lipschitz continuous）。此外，使用对称的JS散度要比不对称的KL散度要好。这个方案认为：训练过程中，判别器过早进入了理想状态，总是很好地识别真实数据，此时如果两个分布距离很远，几乎没有重叠，那么KL散度（KL Divergence）或者JS散度几乎无法给生成器提供梯度信息。因此，即便在两个分布没有重叠的时候，使用Wasserstein Distance（推土机距离）正确地度量两个分布的距离，使得判别器可以更稳定地为生成器提供梯度。

- **谱归一化 Spectral Normalization** 然而求解Wasserstein Distance的计算量较大：通过对权重的奇异值求解，可以得到这一层网络的谱范数（spectral norm），接着让每一层网络的权重除以这一层网络的谱范数就可以满足1-Lipschitz continuous。我们可以采用幂函数迭代法（power iteration）近似地求解谱范数。

- **小批次判别 Mini-batch Discrimination** 使用小批次的数据对判别器进行训练时，生成数据的熵会得到增加。

- **多判别器 Multi-discriminator** 增加判别器的数量，然后使用联合误差对生成器进行训练。

参考文献