디퓨전 모델의 소량 데이터 학습자로서 활용한 다양한 데모 데이터로부터의 오프라인 모사 학습

최윤선°, 황두환, 김기응
한국과학기술원
{cys9506, howdydoo0711, kekim}@kaist.ac.kr

Diffusion model as Few-shot Learner for Offline Imitation Learning from Suboptimal Demonstrations

요 약

강화학습은 다양한 분야에서 환경과의 상호작용을 통해 최적의 행동 정책을 찾아내는 방법으로 활용되고 있다. 그러나 실제 환경에서는 환경과의 실시간 상호작용이 제한되며 에이전트의 행동에 대한 보상 정보의 지속적인 제공이 어렵다. 본 연구에서는 환경과의 상호작용이 제한되고 전문가의 행동 데이터가 소량으로 제공되는 상황 속에서, 전문가의 행동을 효과적으로 모사할 수 있는 모델을 구현하는 것을 목표로 한다. 이를 위해 본 논문에서는 생성모델 중 디퓨전 모델을 활용하며, 배치 데이터 내의 모든 행동을 생성하면서도 샘플링 과정에서 특정 가이던스를 제공하여 전문가 행동을 효과적으로 모방할 수 있도록 모델을 설계하였다.

1. 서론

최근 강화학습 (Reinforcement Learning)은 게임, 바둑, 냉각 시스템 등 다양한 분야에서 성공을 거두며 주목받고 있다 [1, 2]. 강화학습은 환경과의 상호작용를 통해 얻은 정보를 기반으로 누적 보상의 합을 최대화할 수 있는 정책을 학습하기 위한 알고리즘이다. 성공적인 학습을 위해, 온라인으로 상호작용이 가능한 환경과 각 행동을 평가한 보상값을 알아야 하지만, 현실적으로 항상 이를 확보하는 것을 어려운 문제이다. 본 연구에서는 환경과의 상호작용이 불가능한 상황에서 다양한 행동 정책으로 모아진 배치 데이터를 통해 의사결정을 학습하는 것을 목표로 한다. 또한, 각 행동마다 보상값에 대한 정보는 주어지지 않고 적은 갯수의 전문가 행동만 주어져있다. 따라서, 에이전트는 주어진 배치 데이터만을 활용하여 전문가 행동을 효과적으로 모사해야한다.

기존의 방법은 전문가 행동을 바탕으로 보상함수를 학습하여 배치 데이터의 각 행동에 따른 보상값을 라벨링하고 강화학습을 수행하거나 [3], 전문가와 비전문가 행동을 구분하는 판독자를 학습하여 판독자의 결과를 가중치로 사용하여 행동을 모사한다 [4]. 하지만, 이러한 방법은 적은 갯수의 전문가 행동이 주어질 때마다 의사행동을 결정짓는 정책 네트워크와 보상함수 또는 판독자를 다시 학습해야하는 번거로움이 있다.

이러한 문제를 해결하기 위해, 본 논문에서는 디퓨전 (Diffusion) 모델과 비지도 학습 방법으로 배운 latent 변수를 활용한다. 인코더를 학습하여 샘플링하는 과정에서 주어진 적은 갯수의 전문가의 데모와 유사하게 가이던스 [5]를 제공하여, 전문가의 데모가 적게 주어지더라도 효과적으로 모사하는 방법을 제안한다.[1]

1) 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2022-0-00311, 일상적 물건들의 다접촉 로봇 조작을 위한 목적지향 강화학습 기술 개발, No.2020-0-00940, 안전한 강화학습 원천 기술 개발 및 자연어 처리에의 응용)

2. 연구 배경

본 논문에서 다루고자 하는 데이터는 상태 s_t와 행동 a_t의 쌍으로 이루어진 시퀀스, 즉 경로 (trajectory)이다. 이를 나타내면 다음과 같으며, $\boldsymbol{\tau} = \{(s_t, a_t)\}_{t=0}^{T}$, T는 고정된 값이다. 우리는 오프라인 상황에서 다양한 시퀀스로 이루어진 많은 양의 배치 데이터 $\mathcal{D} = \{\boldsymbol{\tau}_j\}_{j=1}^{L}$와 전문가의 행동 시퀀스로 이루어진 적은 갯수의 데모 $\mathcal{D}^{\text{exp}} = \{\boldsymbol{\tau}_j^{\text{exp}}\}_{j=1}^{M}$ $(M << L)$가 주어진 상황에서 효과적인 행동 시퀀스를 생성하는 것을 목표로 한다. 배치 데이터를 생성하는 데이터 분포의 영역은 전문가 행동 분포의 영역을 항상 포함한다고 가정한다.우리는 최근 우수한 표현력으로 주목받는 디퓨젼을 모델을 이용하여, 배치 데이터 $\mathcal{D}$에 존재하는 다양한 행동을 생성하는 디퓨젼 모델을 학습할 것이다.

디퓨젼 모델 디퓨전 모델 (Diffusion probabilistic models) [6, 7]은 반복적인 디노이징 (denoising), $p_\theta(\boldsymbol{\tau}^{i-1}|\boldsymbol{\tau}^i)$, 과정을 통해 데이터를 생성한다. 이 디퓨전 모델은 데이터에 점진적으로 노이즈를 추가하는 순방향 디퓨젼 $q(\boldsymbol{\tau}^i|\boldsymbol{\tau}^{i-1})$ 과정의 역과정이다. 데이터 분포 $p_\theta(\boldsymbol{\tau}^0)$는 다음과 같이 유도한다.

$$p_\theta(\boldsymbol{\tau}^0) = \int p(\boldsymbol{\tau}^N) \prod_{i=1}^{N} p_\theta(\boldsymbol{\tau}^{i-1}|\boldsymbol{\tau}^i) d\boldsymbol{\tau}^{1:N}$$

$p(\boldsymbol{\tau^N})$는 표준 가우시안 사전 (prior) 분포이며, $\boldsymbol{\tau}^0$ 데이터를 나타낸다. 매개변수 θ는 로그-우도 (log-likelihood)의 변분 상한 (variational upper bound)을 최대화하는 방향으로 학습한다.

$$\theta = \arg\max_\theta \mathbb{E}_{\boldsymbol{\tau}^0 \sim \mathcal{D}}[\log p_\theta(\boldsymbol{\tau}^0)]$$

순방향향 디퓨전 과정 $p(\boldsymbol{\tau}^i|\boldsymbol{\tau}^{i-1})$은 일반적으로 사전에 정

의되며, 역방형 디퓨전 과정은 각 디퓨전 스텝마다 다음과 같이 가우시안 분포로 매개변수화된다; $p_\theta(\boldsymbol{\tau}^{i-1}|\boldsymbol{\tau}^i) = \mathcal{N}(\boldsymbol{\tau}^{i-1}|\mu_\theta(\boldsymbol{\tau}^i, i), \Sigma^i)$.

$$p_\theta(\boldsymbol{\tau}^{i-1}|\boldsymbol{\tau}^i) = \mathcal{N}(\boldsymbol{\tau}^{i-1}|\mu_\theta(\boldsymbol{\tau}^i, i), \Sigma^i)$$

표기 본 논문에서는 두 가지 시간 스텝이 있다. 디퓨전 과정에서 순방형, 역방향을 나타내기 위한 시간 스텝과 행동을 취하는 시간 스텝이다. 디퓨전 과정의 시간 스텝을 나타내기 위해서는 윗 첨자 ($\boldsymbol{\tau}^i$)를, 행동을 위한 시간 스텝은 아랫 첨자 (a_t)를 사용한다. 또한 $\boldsymbol{\tau}^i_{s_0}$은 $\boldsymbol{\tau}^i$의 초기 상태를 뜻한다.

3. 본론

오프라인 상황에서 특정 전문가에 대한 데모 데이터가 적게 주어진 상황의 모사 학습을 위해서는 다양한 행동 시퀀드를 생성하는 디퓨젼 모델이 전문가 행동과 유사한 행동을 생성하도록 유도되어야 한다. 이를 위해, 비지도 방식으로 배치 데이터에 존재하는 각 시퀀스 $\boldsymbol{\tau}$를 대표하는 latent 변수 z와 $\boldsymbol{\tau}$를 latent 변수 z로 매핑하는 인코더 $e_\phi(z|\boldsymbol{\tau})$를 학습한다. 학습한 인코더를 이용하여, 전문가 행동 시퀀스 $\boldsymbol{\tau}^{\text{exp}}$가 나타내는 z_{exp}를 인코더를 통해 구한 후, 아래와 같이 인코더와 디퓨션 모델을 이용하면 전문가 행동가 유사한 행동을 생성하는 분포 $p(\boldsymbol{\tau}|z_{\text{exp}})$를 정의할 수 있다.

$$p(\boldsymbol{\tau}|z_{\text{exp}}) \propto p(\boldsymbol{\tau})p(\boldsymbol{\tau}|z_{\text{exp}}) \tag{1}$$

비지도 인코더 학습 행동 시퀀스 $\boldsymbol{\tau}$를 나타낼 수 있는 latent 변수 z를 학습하기 위해 우리는 아래식을 이용하여, 인코더 e_ϕ와 디코더 d_ψ를 학습한다.

$$\min_{\phi,\psi} \mathcal{L}_{\phi,\psi} = -\mathbb{E}_{\boldsymbol{\tau}\sim\mathcal{D}}[\mathbb{E}_{z\sim e_\phi(z|\boldsymbol{\tau})}[\sum_{t=0}^{T} \log d_\psi(a_t|s_t, z)]$$

인코더 매개 변수 ϕ은 시퀀스를 처리하기 위한 GRU의 웨이트로 구성되며, 가우시안 분포를 매개화한다; $e_\phi(z|\boldsymbol{\tau}) = \mathcal{N}(z|GRU_\phi(\boldsymbol{\tau}), \sigma_z I)$.

$$e_\phi(z|\boldsymbol{\tau}) = \mathcal{N}(z|GRU_\phi(\boldsymbol{\tau}), \sigma_z I)$$

$$d_\psi(a_t|s_t, z) = \mathcal{N}(a_t|\mu_\psi(s_t, z), \sigma_a I)$$

디코더 매개 변수 ψ는 현재 상태가 주어졌을 때 행동을 결정하는 가우시안 분포를 매개화한다; $d_\psi(a_t|s_t, z) = \mathcal{N}(a_t|\mu_\psi(s_t, z), \sigma_a I)$. 여기서 σ_z, σ_a는 각 랜덤 변수의 표준편차를 뜻하며, 본 논문에서는 고정된 값을 이용하였다.

Algorithm 1 Guided Diffusion Sampling

Require : Diffusion model μ_θ, covariance Σ^i, scale α, encoder e_ϕ, a few expert demonstrations $\mathcal{D}^{\text{exp}} = \{\boldsymbol{\tau}^{\text{exp}}_j\}_{j=1}^M$
Observe initial state s_0; initialize $\boldsymbol{\tau}^N \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
Sample $z_{\text{exp}} \sim e_\phi(\cdot|\mathcal{D}^{\text{exp}})$ in Eqn. (2)
for $i = N, ..., 1$ **do**
  $\mu \leftarrow \mu_\theta(\boldsymbol{\tau}^i)$
  // guide using gradients of e_ϕ
  $\boldsymbol{\tau}^{i-1} \sim \mathcal{N}(\mu + \alpha\nabla_{\boldsymbol{\tau}=\mu} \log e_\phi(z_{\text{exp}}|\boldsymbol{\tau}), \Sigma^i)$
  // inpainting $\boldsymbol{\tau}^{i-1}_{s_0}$ as s_0
  $\boldsymbol{\tau}^{i-1}_{s_0} \leftarrow s_0$
end for

특정 전문가의 소량의 데모를 이용한 Classifier-guided 샘플링 전문가의 행동 데모는 하나 이상으로 주어진다. 주어진 모든 전문가 행동 시퀀스 $\mathcal{D}^{\text{exp}} = \{\boldsymbol{\tau}^{\text{exp}}_j\}_{j=1}^M$를 고려한 latent 변수 z^{exp}를 결정하기 위해, 순열에 불변한 곱셉 형식 (Product-of-experts)의 모델링을 아래와 같이 이용하였다.

$$z_{\text{exp}} \sim e_\phi(\cdot|\mathcal{D}^{\text{exp}}) \propto \prod_{\boldsymbol{\tau}^{\text{exp}}\in\mathcal{D}^{\text{exp}}} e_\phi(z|\boldsymbol{\tau}^{\text{exp}}) \tag{2}$$

$e_\phi(\cdot|\mathcal{D}^{\text{exp}})$ 또한 가우시안 분포이며, 이 분포의 평균와 표준편차는 각 요소 분포 $e_\phi(z|\boldsymbol{\tau}^{\text{exp}})$의 평균과 표준편차로 결정된다.

이렇게 결정한 가우시언 분표에서 z^{exp}를 샘플링한다. 샘플링된 latent 변수를 z_{exp} 이용하여, 로그-확률을 높이는 경사방향 g를 디퓨젼 모델에게 가이던스로 제공할 수 있다.

$$g = \nabla_{\boldsymbol{\tau}} \log e_\phi(z_{\text{exp}}|\boldsymbol{\tau}) \tag{3}$$

식 (3)의 가이던스와 배치 데이터로 학습한 디퓨젼 모델을 이용하여 알고리즘 1로 행동 시퀀스를 생성하면, 전문가 행동이 주어졌을 때마다 그와 유사한 행동을 생성할 수 있다.

4. 실험

적은 수의 특정 전문가 데모가 주어질 때마다 제안한 알고리즘이 전문가 데모와 유사한 행동을 생성하는지를 평가하기 위해 Point robot 환경에서 실험을 수행하였다. 구체적으로 1개부터 16개의 전문가 데모 ($M = 1, 4, 8, 16$)가 주어졌을 때, 디퓨젼 모델이 알고리즘 1을 이용하여 생성하는 행동 시퀀스를 확인하였다.

실험 세팅 Point robot 환경은 로봇이 초기 위치 (0, 0)에서 시작하여 지정된 목표에 도달하는 것을 목표로 한다. 목표는 (0, 0)을 중심으로 한 지름 1의 원 위에 위치한다. 디퓨젼 모델을 학습하기 위해, 무작위로 3000개의 목표을 샘플링하고 이 목표에 도달하는 3000개의 경로를 배치 데이터로 사용하였다. 경로 길이는 $T = 64$로 가이던스 스케일은 $\alpha = 1$로 설정하여 실험을 진행하

였다.

결과 그림 1은 전문가 데모가 주어졌을 때 생성하는 행동을 (x, y) 위치 좌표의 시퀀스로 보여주고 있다. 총 4가지 목표에 대해 1, 4, 8, 16개의 전문가 데모를 이용하였다. 그림 1의 왼쪽 열에서는 전문가의 데모를 나타내며, 동일한 색상의 전문가 경로가 z_{exp}를 추론하기 위해 사용되었다. 구체적으로 그림 1 (a)에서는 각 목표에 대해 하나의 데모가 z_{exp}를 결정하기 위해 사용되었으며, (g)에서는 16개의 전문가 데모가 사용되었다. 그림의 오른쪽 열에서는 결정된 z_{exp}를 바탕으로 생성된 16개의 예시 행동 시퀀스를 보여준다. 이를 통해, 본 실험에서 전문가의 데모와 비슷한 행동 시퀀스를 생성함을 확인할 수 있다. 또한 전문가의 데모가 새롭게 주어지더라도 제안한 가이디드를 통해 샘플링하면 모델을 다시 학습할 필요없이 이와 같은 결과를 얻을 수 있음을 확인하였다.

5. 결론

본 연구에서는 전문가 행동 데이터가 제한적으로 제공되는 오프라인 환경에서 효과적인 행동 모사를 위한 디퓨전 모델의 활용과 관련된 방법을 탐구하였다. 본 연구에서는 이 문제에 대한 해결책으로 디퓨전 모델과 비지도 학습을 통해 구한 latent변수를 활용하였다. 디퓨전 모델은 반복적인 디노이징 과정을 통해 다양한 행동 시퀀스를 생성하는데 사용이 되었으며, 배치 데이터내의 각 시퀀스를 대표하는 latent 변수를 학습하고 시퀀스를 latent 변수로 매핑할 수 있는 인코더를 학습하였다. 이를 통해 전문가 행동 시퀀스를 인코딩하여 전문가 행동과 유사한 행동을 생성하는 분포를 정의할 수 있었다. 실험을 통해 이 방법이 전문가 데모의 수와 관계없이 유사한 행동 시퀀스를 생성해 낼 수 있음을 확인하였으며, 새로운 데이터가 주어지더라도 본 논문에서 제안한 가이던스를 통해 샘플링을 진행할 경우 재학습에 대한 요구없이도 동일한 결과를 도출할 수 있음을 확인하였다.

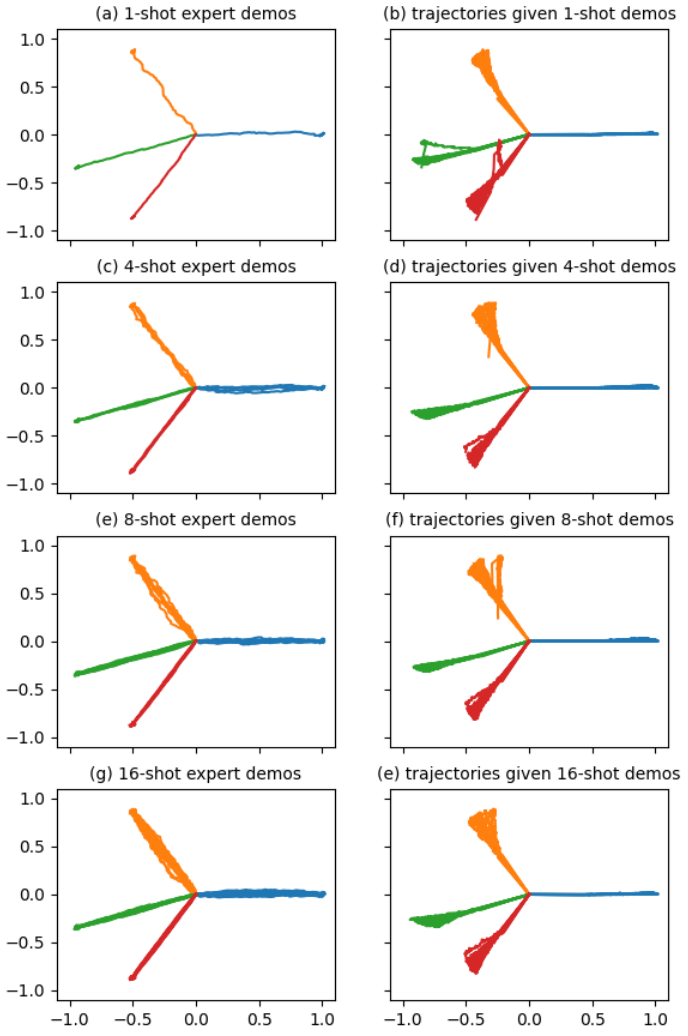

그림 1: 실험 결과. 왼쪽 열은 4개의 골에 도달하기 위해 주어진 전문가 행동을 나타냄. 오른쪽 열은 각 골마다 주어진 전문가의 행동을 바탕으로 디퓨전 모델이 생성한 행동을 보여줌. 각 열 마다 1, 4, 8, 16 갯수의 전문가 행동이 주어짐.

참고 문헌

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," 2013.

[2] D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484–489, 01 2016.

[3] G.-H. Kim, S. Seo, J. Lee, W. Jeon, H. Hwang, H. Yang, and K.-E. Kim, "DemoDICE: Offline imitation learning with supplementary imperfect demonstrations," in *International Conference on Learning Representations*, 2022.

[4] H. Xu, X. Zhan, H. Yin, and H. Qin, "Discriminator-weighted offline imitation learning from suboptimal demonstrations," in *Deep RL Workshop NeurIPS 2021*, 2021.

[5] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan,

and H. Lin, eds.), vol. 33, pp. 6840–6851, Curran Associates, Inc., 2020.

[7] M. Janner, Y. Du, J. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis," in *International Conference on Machine Learning*, 2022.