데이터 기반 심층강화학습을 통한 모션 생성*

박근태[0], 이윤상
한양대학교 컴퓨터소프트웨어공학과
{qkrrmsxo01, yoonsanglee}@hanyang.ac.kr

Motion Generation using Data-driven Deep Reinforcement Learning

Geuntae Park[0], Yoonsang Lee
Dept. of Computer Science, Hanyang University

요약

본 논문에서는 물리 시뮬레이션 없이 GAN 구조를 활용한 강화 학습을 통한 데이터 기반 모션 생성 방법을 제안한다. 네트워크는 주어진 목표를 달성하는 것 외에도 실제 사람의 동작에 기반한 데이터를 통해 자연스러운 모션을 생성하는 방법을 학습한다.

1. 서론

캐릭터 동작 생성 분야에서 많은 심층강화학습을 활용한 연구들은 물리 시뮬레이션을 이용하는 물리 기반 (physics-driven) 접근법을 활용했다 [1, 2]. 물리 기반 접근법은 물리 법칙을 계산하여 동작을 생성하는 접근 방법으로 이를 통해 물리적으로 사실적인 동작을 생성할 수 있다. 하지만 물리적인 사실성을 넘어서 실제 사람과 같이 자연스러운 동작을 유지하며 다양한 과제를 수행하도록 하기 위해서는 추가적인 노력이 필요하며, 물리 법칙 계산에 따른 계산량의 증가도 존재한다.

데이터 기반 (data-driven) 접근법은 실제 사람의 동작을 캡처한 모션 캡처 데이터를 이용하여 동작을 생성한다. 물리 법칙을 계산하여 동작을 생성하는 물리 기반 접근법과는 달리 실제 환경에서 얻은 자연스러운 동작 데이터를 바로 사용하므로 물리 법칙을 위한 별도의 계산이 필요하지 않으며, 생성되는 동작 역시 실제 사람처럼 자연스럽다는 장점이 있다. 본 논문에서는 데이터 기반 접근법에 기반한, 심층강화학습을 사용한 모션 생성 방법을 제안한다.

우리가 제안하는 시스템은 현재 프레임의 모션을 생성하는 생성자 (generator) 네트워크와, 일련의 모션 데이터가 실제 모션 캡처 데이터인지 아니면 생성자에 의해 만들어진 데이터인지 판별하는 판별자 (discriminator) 네트워크로 이루어진 generative adversarial network (GAN) [3] 구조로 이루어져 있다. 생성자는 목표를 달성하는 동작으로 생성하도록 강화학습을 통해 학습되는 정책 (policy)이다. 실제 모션 데이터에 가까운 자연스러운 동작을 생성하도록 학습 과정에서 판별자의 출력 값이 보상 (reward)에 반영된다.

* 구두 발표논문, 요약논문 (Extended Abstract)
* 본 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행되었음 (NRF-2019R1C1C1006778).

2. 방법

2.1. 네트워크 구조

전체 네트워크는 크게 생성자와 판별자로 구성된 GAN의 형태로 구성된다. 여기서 생성자는 강화 학습을 통해 정책을 학습하는 강화 학습 네트워크이고 판별자는 생성자의 출력 값과 실제 데이터를 구별하도록 학습되는 간단한 형태의 Fully Connected Network이다. 실제 동작에는 학습이 완료된 생성자의 출력 값만 사용된다.

2.2. 생성자

생성자는 proximal-policy optimization (PPO) [4]를 통해 정책을 학습한다. 각 시간 t에서 에이전트 (agent)는 목표 보상 (goal-reward)과 판별 보상 (discriminator-reward), 총 2 종류의 보상을 받는다. 목표 보상은 캐릭터가 수행하기를 원하는 과제에 맞춰 임의로 설정할 수 있다. 판별 보상은 t 시점에서 에이전트의 동작 정보를 판별자에 입력하여 얻은 판별 값을 통해 계산된다.

$$\boldsymbol{r} = \boldsymbol{w_g r_g} + \boldsymbol{w_d r_d} \quad (1)$$

위의 식에서 w_g는 최종 보상 r 에 반영할 목표 보상 r_g의 가중치이며 w_d는 판별 보상 r_d의 가중치이다. (r_g와 r_d에 관한 자세한 수식은 각각 3 장과 2.3 절 참조).

생성자는 목표 보상을 통해 현재 상황에서 주어진 목표를 달성하는 방법을 학습하며 판별 보상을 통해 그 과정에서 자연스러운 동작을 만드는 방법을 학습한다. 학습이 완료된 생성자는 다음 프레임에서의 동작 정보와 위치 변화량 (offset)을 출력한다.

2.3. 판별자

판별자는 모션 캡처 데이터를 학습에 사용하는 감독 학습 (supervised learning)을 수행하는 3 계층 (layer) 네트워크로 구성되었다. 기존의 GAN은 판별자에 sigmoid cross-entropy를 사용한다. 그러나 sigmoid는 vanishing gradient가 발생하는 문제가 있다. 이를 방지하기 위해 본 논문에서는 sigmoid cross-entropy 대신 least square를 사용하는 least-squares GAN (LSGAN) [5]의 판별자를 사용했다. 판별자 학습에는 다음과 같은 손실함수가 사용되며 이를 통해 판별자는 [-1, 1] 범위 내의 값으로 입력 받은 데이터의 진위를 표현한다.

$$\arg\min_{D} \mathbb{E}_{d\in R}\left[(D(d)-1)^2\right] + \mathbb{E}_{d\in G}\left[(D(d)+1)^2\right] \quad (2)$$

위 식에서 D는 판별자를 나타내며 R은 실제 데이터셋을, G는 생성자를 나타낸다.

판별자는 이전 프레임과 현재 프레임에서의 캐릭터 동작 정보 s_{t-1}, s_t를 입력으로 받는다. 캐릭터의 동작 정보는 각 관절의 3차원 회전 벡터로 표현했다.

$$r_d = \max[0, 1 - 0.25(D(s_{t-1}, s_t) - 1)^2] \quad (3)$$

식 (3)은 식(1)에서 서술한 판별 보상 식으로 [0, 1] 사이의 값을 얻기 위해 판별자 출력 값에 추가적인 계산을 수행했다.

3. 실험 결과

생성자와 판별자의 학습 방식 차이로 두 네트워크 간 학습 속도의 차이가 발생하는 것을 발견했다. 이에 실험 환경에서는 학습 속도에 균형을 맞추고자 생성자와 판별자의 학습 횟수를 100 대 1의 비율로 설정했다. 실험을 위해 약 30시간 정도 학습을 진행했다.

실험 환경에서 캐릭터의 목표는 원점에서 출발하여 임의의 장소 x_g에 도달하는 것으로 설정했다. 학습에 사용된 목표 보상 식은 두 지점 사이의 거리로 계산되며 다음과 같이 표현된다.

$$r_g = \exp(-0.5\|x_g - x_t\|^2) \quad (4)$$

위 식에서 x_g는 목표 지점을 나타내며 x_t는 시간 t에서의 캐릭터의 위치를 나타낸다. x_g와 x_t는 전역 좌표계로 표현된 지면상의 위치를 사용한다. 두 보상 r_g과 r_d의 비중을 동일하게 만들기 위해 식(1)의 가중치 w_g와 w_d 모두 0.5로 설정했다.

판별자의 학습에는 두 지점을 여러 차례 왕복하는 동작을 30 fps으로 기록한 약 8000 프레임의 길이를 가진 단일 모션 데이터를 통해 학습을 진행했다.

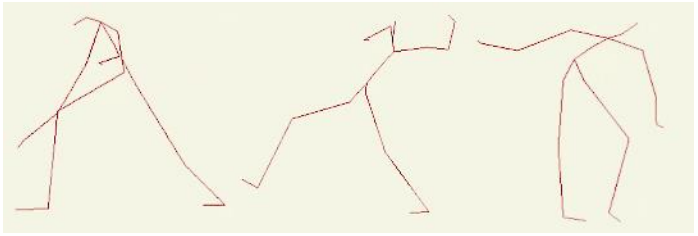

그림 1: 생성자 출력 모션 예시

실험 환경에서 목표는 크게 목적지에 도달하는 것과 자연스러운 동작 생성 두 가지로 이루어져 있다. 결과물을 통해 첫 번째 목표는 어느 정도 달성하는 것을 확인했다. 하지만 두 번째 목표의 경우 개별적으로는 사람이 해 낼 수 있는 동작을 보였으나 자연스럽게 전환이 이루어지지 않아 제대로 달성하지 못하는 것을 확인했다 (그림 1 참조). 또한 학습 횟수의 비율을 달리 했음에도 불구하고 생성자가 판별자에 비해 느리게 학습이 되는 현상이 나타났는데 이는 학습이 까다로운 GAN 기반 모델의 특성으로 인한 것으로 보이며 비율 조정 외의 추가적인 학습 성능 향상을 위한 기법들을 적용해야 할 필요가 있을 것으로 생각된다.

4. 결론

강화 학습에 모션 캡처 데이터를 통해 학습된 판별자를 이용하는 방법으로 데이터 기반 접근법의 적용 가능성을 확인했다. 하지만 감독 학습을 수행하는 판별자의 학습 속도를 생성자가 따라잡지 못하는 문제가 해결되지 않아 생성된 모션의 품질이 떨어지는 현상이 확인되었다. 앞으로의 연구는 생성자와 판별자 사이의 학습 속도 차이를 개선하는 방향으로 진행될 것이다.

참고문헌

[1] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. 2018a. Deep-Mimic: Example-guided Deep Reinforcement Learning of Physics-based Character Skills. *ACM Trans. Graph.* 37, 4, Article 143 (July 2018), 14 pages.

[2] Zhiqi Yin, Zeshi Yang, Michiel Van De Panne, and KangKang Yin. Discovering diverse athletic jumping strategies. *ACM Transactions on Graphics (TOG)*, 2021.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, DavidWarde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2672–2680.

[4] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347 (2017). arXiv:1707.06347

[5] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley. 2017. Least Squares Generative Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2813–2821.